From image coding and representation to robotic vision

Marie Babel

Habilitation à Diriger des Recherches

From image coding and representation to robotic vision

Marie BABEL

Université de Rennes 1
June 29th 2012

Bruno Arnaldi, Professor, INSA Rennes	Committee chairman
Ferran Marques, Professor, Technical University of Catalonia	Reviewer
Benoît Macq, Professor, Université Catholique de Louvain	Reviewer
Frédéric Dufaux, CNRS Research Director, Telecom ParisTech	Reviewer
Charly Poulliat, Professor, INP-ENSEEIHT Toulouse	Examiner
Claude Labit, Inria Research Director, Inria Rennes	Examiner
Françis Chaumette, Inria Research Director, Inria Rennes	Examiner
Joseph Ronsin, Professor, INSA Rennes	Examiner

IRISA UMR CNRS 6074 / INRIA - Equipe Lagadic
IETR UMR CNRS 6164 - Equipe Images

Contents

Remerciements

Cette habilitation est une façon pour moi de faire un bilan sur mes activités passées au sein du laboratoire IETR et mettre en exergue le lien qui existent avec celles qui se déroulent désormais à l'IRISA, dans l'équipe Lagadic.

Rien n'est linéaire, ni les relations humaines, et encore moins le cheminement de la pensée. Si des chemins de traverse sont pris, c'est pour mieux renforcer les convictions et la détermination que l'on a à accomplir ce pour quoi on se bat.

Je tiens à remercier très sincèrement tout d'abord tous les membres de mon jury pour leur participation à cette entreprise particulière qu'est l'Habilitation à Diriger les Recherches. Merci infiniment à mes rapporteurs Ferran Marquès, Benoît Macq et Frédéric Dufaux pour leur relecture attentive et leur soutien. Merci infiniment aussi à Charly Poulliat, Denis Friboulet, Claude Labit pour les échanges conviviaux et très riches. Un grand merci à Joseph Ronsin, mon directeur de thèse à l'IETR, pour lequel j'ai un profond respect.

Il me semble aussi important de remercier tous ceux qui ont facilité mon intégration à l'IRISA. En tout premier, je remercie Bruno Arnaldi pour l'aide précieuse, l'attention qu'il m'a témoignée et bien sûr la présidence de cette habilitation. François Chaumette m'a alors accueillie avec chaleur : je lui en suis vraiment reconnaissante. Laurent Bédat, Muriel Pressigout, Sylvain Guégan ont toujours été à mes côtés dans les moments les plus difficiles et leur confiance me touche.

Côté Lagadic, Alexandre Krupa, Fabien Spindler, Eric Marchand doivent être chaleureusement remerciés eux aussi pour leur attention sincère à mon égard et l'accueil qu'ils m'ont réservé. Sans cet entourage fort, je n'aurai certainement pas été à même de me projeter dans le futur et mener à bien une mobilité thématique.

Ivan Leplumey ne peut pas être oublié dans ces remerciements. Il m'a invitée avec une infinie gentillesse à prendre place au sein département Informatique de l'INSA. La convivialité, la solidarité ne sont pas des vains mots...

Marie Françoise Pichot, Nathalie Jacquinot, Céline Gharsalli sont des assistantes dont les qualités humaines et le dévouement auprès des équipes de recherche ou pédagogiques n'ont d'égal que ceux de Jocelyne Trémier, pour laquelle mon amitié restera indéfectible.

Merci aussi à tous ceux à l'INSA avec lesquels je partage des aventures pédagogiques et quotidiennes dans la bonne humeur : Jean-Noël Provost, Luce Morin, Sylvain Haese, Eric Bazin, Maxime Pelcat, Véronique Coat, Kidiyo Kpalma, Jean-Gabriel Cousin, Laurence Rozé, Marin Bertier...

Je ne peux finir cette liste à la Prévert sans remercier du fond du coeur mes petits canetons (autre dénomination d'un doctorant) : François Pasteau que j'ai adopté définitivement, Clément Strauss, Fabien Racapé, Rafik Sekkal, Matthieu Urvoy, Yohann Pitrey, Jean Motsch. Sans vous, cette habilitation n'aurait pu être réalisée... Vous êtes épatants !

Je dédie cette HDR à mon père.
A Aziliz, Klervi et Julien.

Chapter 1

Introduction

1.1 An overview of my research project

This habilitation thesis is somehow specific since the main part of my work unfolded within the IETR laboratory although I recently joined the IRISA / INRIA laboratory. This situation reflects the natural evolution of my research topics.

My research works were in fact first oriented towards image and video coding, and relied on the Locally Adaptive Resolution (LAR) coding scheme. Then, image representation and analysis domains were introduced in order to fit human vision properties as well as the image content. The combination of these topics led us to look into joint analysis and coding frameworks.

Meanwhile, as the current French National Research Agency (ANR) encourages the researchers to take part in and to manage projects for both collaboration and technological transfer purposes, more and more research works are directly driven by project purposes. Clearly, our work has been naturally influenced by the different collaborative projects in which we were involved. As a consequence, my research topics are not exclusive and provide a large range of analysis and coding solutions. Even if they are strongly oriented towards compression, I took an interest in a large panels of related topics.

In this way, I participated in two ANR projects, both of them addressing securization issues. The TSAR project (Safe Transfer of high Resolution Art images) tended to design a scalable solution of securized transmission of coded images. As for it, the CAIMAN project (Codage Avancé d'IMAges et Nouveaux services) aims at designing a new joint source/channel coding framework able to integrate Human Vision System based features. Hence, I worked on data integrity issues through dedicated cryptography and steganography solutions and through Quality of Services oriented techniques. In addition, the main objective of CAIMAN project was to propose a coding solution as a response to the JPEG-AIC call for proposal. As the LAR was chosen to be a candidate, a large part of the team's work was devoted to an active participation to JPEG committee actions.

In connection to these research fields, since 2005, I co-supervised with Joseph Ronsin and Olivier Déforges six PhD students that have defended their dissertations, namely Jean Motsch, Yohann Pitrey-Helpiquet, Matthieu Urvoy, François Pasteau, Fabien Racapé and Clément Strauss.

Recently, researches have been even more focused on pseudo-semantic representation of images and videos. If these studies have been conducted so that to match compression issues, related domains can be investigated. This leads me to tackle robotic vision issues, and more particularly robotic assistance. I am now involved into different projects that clearly influence my future research.

I obtained, by special dispensation, the authorization to independently supervise two PhD students (Rafiq Sekkal, Yao Zhigang), whose works are oriented towards tracking and autonomous navigation with the help of 3D vision systems. For all these reasons, I joined the Lagadic Team within IRISA / INRIA laboratory in November 2011.

1.2 Coding and representation tools: QoS/QoE context

This habilitation document is then mainly devoted to applications related to image representation and coding. Of course, this research field has been deeply investigated and many related previous books [205] and habilitation thesis have been already published [146][38][129] in this domain. The idea is not here to rewrite the image coding context, but rather to reinstate this work among emerging services.

If the image and video coding community has been traditionally focused on coding standardization processes, advanced services and functionalities have been designed in particular to match content delivery system requirements. In this sense, the complete transmission chain of encoded images has now to be considered.

To characterize the ability of any communication network to insure end-to-end quality, the notion of Quality of Service (QoS) has been introduced. First defined by the ITU-T as the set of technologies aiming at the degree of satisfaction of a user of the service [85], QoS is rather now restricted to solutions designed for monitoring and improving network performance parameters. However, end users are usually not bothered by pure technical performances but are more concerned about their ability to experience the desired content. In fact, QoS addresses network quality issues and provides indicators such as jittering, bandwidth, loss rate...

An emerging research area is then focused on the notion of Quality of Experience (QoE, also abbreviated as QoX), that describes the quality perceived by end users. According to [86], QoE is defined as a subjective metric measuring the acceptability of an application of a service. Within this context, QoE faces the challenge of predicting the behaviour of any end users.

When considering encoded images, many technical solutions can considerably enhance the end user experience, both in terms of services and functionalities, as well as in terms of final image quality. Ensuring the effective transport of data, maintaining security while obtaining the desired end quality remain key issues for video coding and streaming [162].

First parts of my work are then to be seen within this joint QoS/QoE context. From efficient coding frameworks, additional generic functionalities and services such as scalability, advanced entropy coders, content protection, error resilience, image quality enhancement have been proposed.

1.3 Image and video representation: towards pseudo-semantic technologies

Related to advanced QoE services, such as Region of Interest definition of object tracking and recognition, we further closely studied pseudo-semantic representation. First designed toward coding purposes, these representations aim at exploiting textural spatial redundancies at region level.

Indeed, research, for the past 30 years, provided numerous decorrelation tools that reduce the amount of redundancies across both spatial and temporal dimensions in image sequences. To this

day, the classical video compression paradigm locally splits the images into blocks of pixels, and processes the temporal axis on a frame by frame basis, without any obvious continuity. Despite very high compression performances such as AVC [206] and forthcoming HEVC standards [208], one may still advocate the use of alternative approaches. Disruptive solutions have also been proposed, and offer notably the ability to continuously process the temporal axis. However, they often rely on complex tools (*e.g.* Wavelets, control grids [29]) whose use is rather delicate in practice.

We then investigate the viability of alternative representations that embed features of both classical and disruptive approaches. The objective is to exhibit the temporal persistence of the textural information, through a time-continuous description.

At last, from this pseudo-semantic level of representation, texture tracking system up to object tracking can be designed. From this technical solution, 3D object tracking is a logical outcome, in particular when considering vision robotic issues.

1.4 Organization of the document

This document is divided into five main parts related to coding, securization and analysis solutions. These contributions are related to my work within the IETR laboratory.

The contributions presented in chapter 2 fit into the schemes of the LAR coding framework. The scalable version of the LAR, namely the Interleaved S+P framework, is particularly described. From this scheme, we designed advanced generic tools, presented in chapter 3, specially defined for any predictive coders. In particular, relying on error prediction distributions studies, generic color-based decorrelation solution as well as as specific entropy coder are described.

Chapter 4 addresses data integrity issues. Scalable cryptography and steganography solutions are first proposed, then network-oriented Quality of Service solutions, elaborated for scalable coder purposes, are shown. Generic analysis tools, that tend to match low complexity requirements, are defined in chapter 5. An original Dyadic Fast Interpolation is thus proposed, along with quadtree-based segmentation solutions. Chapter 6 presents then disruptive approaches for joint representation and coding of image sequences.

The last chapter (chapter 7) states for my research perspectives, oriented towards robotic vision. It emphasizes my progressive transition towards assistance living concerns through adapted technologies.

Chapter 2

Still image coding and advanced services

Nowadays, easy-used communication systems have emphasized the development of various innovative technologies including digital image handling, such as digital cameras, PDAs or mobile phones. This naturally leads to implement image compression systems used for general purposes like digital storage, broadcasting and display. JPEG, JPEG 2000 and now JPEG XR have become international standards for image compression needs, providing efficient solutions at different complexity levels. Nevertheless, if JPEG 2000 is proved to be the most efficient coding scheme, its intrinsic complexity prevents its implementation on embedded systems that are limited in terms of computational capacity and/or memory. In addition, usages associated with image compression systems are evolving, and tend to require more and more advanced functionalities and services that are not always well addressed by current norms. As a consequence, designing an image compression framework still remains a relevant issue.

The JPEG committee has started to work on new technologies to define the next generation of image compression systems. This future standard, named JPEG AIC (Advanced Image Coding), aims at defining a complete coding scheme able to provide advanced functionalities such as lossy to lossless compression, scalability, robustness, error resilience, embed-ability, content description for image handling at object level [98].

In this context, the Locally Adaptive Resolution (LAR) codec framework, designed within the IETR laboratory and initiated by Olivier Déforges, has been proposed as a contribution to the relative call for technologies, tending to fit all of previous functionalities [40]. The related method is a coding solution that simultaneously proposes a relevant representation of the image. The LAR method has been initially introduced for lossy image coding. This original image compression solution relies on a content-based system driven by a specific quadtree representation. Multiresolution versions of this codec have shown their efficiency, especially for lossless coding purposes. An original hierarchical self-extracting region representation has also been elaborated: a segmentation process is realized at both coder and decoder, leading to a free segmentation map. This latter can be further exploited for color region encoding or image handling at region level. Thanks to the modularity of our coding scheme, the complexity can be adjusted to address various embedded systems. For example, basic version of the LAR coder has been implemented onto FPGA platforms while respecting real-time constraints. Pyramidal LAR solution and hierarchical segmentation processes have also been prototyped onto

DSPs heterogeneous architectures.

During my own PhD work, my main contribution was devoted to the design of a scalable coding framework based on the LAR principles. The Interleaved S+P solution is thus based on a two interlaced pyramidal representation and is used for coding purposes [21]. If other versions of a multiresolution LAR coding solution have been studied [41], the Interleaved S+P, thanks to its contained complexity and the advanced functionalities it proposes, is the framework that has been further developed and presented as a response to JPEG-AIC Call For Proposal. Since then, I co-supervised four PhD works relative to this topic.

In this chapter, we focus on the Interleaved S+P still image coding characteristics. JPEG AIC scope is first introduced as well as associated requirements. Then we describe the technical features of the LAR system, and show the originality and the characteristics of the proposed scheme.

2.1 JPEG AIC calls for proposal: a constrained applicative context

A part of my research work has been realized within the ANR CAIMAN project[1]. A major objective of this project was to propose an innovation coding solution that could be able to fit the new JPEG-AIC standard requirements. The LAR image codec has been then selected as a possible answer to dedicated call for proposal. As a consequence, our work has undergone the JPEG expectations in terms of both scientific developments and scheduling. This section presents then our involvement in the JPEG-AIC normalization process.

2.1.1 Evolution of codecs: JPEG committee

The most widely used image codec is JPEG for sure. This codec has been standardized by the JPEG committee in 1992 [78]. It is mainly based on a Discrete Cosine Transform (DCT), together with a dedicated quantization process and Huffman entropy coding. The JPEG codec has been then further improved by the mean of a Q15 binary arithmetic coder [81]. This image codec exhibits interesting characteristics such as a low complexity together with good quality at medium to high bit rates. However it lacks of lossless mode, preventing it from encoding images without any artefact. The JPEG format is widely used in digital cameras as well as on the Internet.

To address the issue of lossless encoding, JPEG-LS has been standardized by the JPEG committee in 1998 [79]. JPEG-LS uses a completely different approach than JPEG does and is therefore not backward compatible with JPEG. JPEG-LS is a predictive codec based on the LOCO-I scheme [203].

In 2002, JPEG2K has been standardized [80] by the JPEG committee. This image codec is based on a Cohen-Daubechies-Feauveau wavelet. It offers resolution scalability, high compression ratio, rate control and ROI management. However it lacks of JPEG backward compatibility and has a high complexity.

As for it, JPEG Wireless [82] is based on the JPEG 2K image codec. It addresses the issue of transmitting image through error prone wireless network. It offers a new layer of error robustness to the JPEG2K codec.

Standardized in 2010, JPEG-XR [83] created by Microsoft offers a low complexity, average rate distortion [159], complexity [139] results between JPEG and JPEG2K, and lacks of JPEG or JPEG2K

[1] http://www.research-projects.org/xwiki/bin/view/CAIMAN/

backward compatibility.

Within this compression scope, WebP[65], created by Google, is studied by the JPEG committee to check its relevance. It corresponds to the Intra mode of VC8 also known as WebM [64]. It offers good rate distortion performance compared to the state of the art JPEG2K and JPEGXR [49]. However it lacks of any kind of scalability.

To propose a new image codec, the JPEG committee has issued in 2007 a new Call for Advanced Image Coding (AIC) and evaluation methodologies [98]. The committee has then initiated an activity to study potential methodologies for image quality evaluation and technologies for future generation image compression systems. This activity is seen as a new work item for standardization of a new evaluation approaches and image compression system if any potential technologies, which significantly improve the performance current image compression standards, can be identified. Among others, an area of special interest is the creation of a low-complexity high-performance coding format for two, three or higher-dimensional medical data.

2.1.2 Response to the call for JPEG-AIC

The LAR codec has been proposed as a response to the Call for Advanced Image Coding (AIC) and evaluation methodologies [98]. We responded to the call for both natural images (contribution [12]) and for medical images (contributions [19, 20]). In addition we answered to the call for a medical image database (contribution [13]).

Once the response to the AIC call for proposal has been accepted in July 2010, the work was organized around several "core experiments" aiming at testing specific functionalities of the codec. Each core experiment must be successively validated to continue the JPEG process. Therefore the LAR codec has been optimized for the different core experiments while being generally improved in terms of ease of use, parameterization, architecture, complexity, modularity, and while making sure that everything remains bug free. Four versions of the reference software have been published all along this process [14, 15, 16, 17].

This implication in JPEG committee deeply influenced the team's work. Unfortunately this software development, testing process and in general the JPEG contributions around the LAR codec do not usually results in promotable scientific contributions while being very time consuming together with imposing some constraints. However the visibility of our work has been greatly increased and has lead to interesting exchanges with the international image coding community. The PhD works of both François Pasteau and Clément Strauss were conducted within this context. Together with Médéric Blestel, an expert engineer specially recruited for the CAIMAN project, they provide major innovations towards JPEG-AIC requirements.

The following paragraphs present 3 core experiments and summarize the corresponding contributions.

Core 1: objective quality, July 2010 - October 2010. This core experiment [99], realized from July 2010 up to October 2010, was targeting on objective quality. The core results showed some problems with the codec due to parametrization process, as ideal parameters needed to be found for each image. Testers used a standard set of parameters leading to some sub-optimal results. However a good behavior of the codec with respect to the JPEG standards has been emphasized.

Core 2: objective and subjective quality, October 2010 - February 2011. The second core [100], from October 2010 up to February 2011, added subjective quality measures into the tests. The LAR reference software [16] has been evaluated in a subjective manner by the university of Poitiers [101] and in a objective manner by the university of Stuttgart [160]. The result of the core showed that the LAR is performing similarly to JPEG-XR for 0.25 and 0.5bpp, and was worst for higher bitrates. The conclusion was that some improvements were necessary to help find the LAR optimal parameters.

Core 3: performance evaluation and functionality analysis, February 2011 - July 2011. For the third core [102], from February 2011 up to July 2011, the LAR reference software [17] has been evaluated in terms of complexity and objective quality [18]. We conducted and presented both evaluations. The results of the core, presented in July 2011, showed a lower performance in terms of complexity and compression when compared to JPEG2K. As a consequence the JPEG committee decided to abandon the experiments on the LAR codec. However this participation remains a very positive experience, especially in terms of both industrial contacts and normalization process understanding.

2.2 Locally Adaptive Resolution compression framework: an overview

The LAR method, first designed by Olivier Déforges in [42], was initially introduced for lossy image coding purposes [40]. The philosophy behind this coder is not to outperform JPEG2K in terms of compression, but rather to propose an open source, royalty free, alternative image coder with integrated services. While keeping the compression performances in the same range as JPEG2K(lossless coding) or JPEG XR (lossy coding), with contained complexity, our coder also provides services such as scalability, cryptography, data hiding, free region representation and coding.

The LAR codec is based on the assumption that an image can be represented as layers of basic information and local texture, relying then on a two-layer system (figure 2.1). The first layer, called Flat coder, leads to construct a low bit-rate version of the image with good visual properties. The second layer deals with the texture that is encoded through a dedicated texture coder relying on DCT (spectral coder) or pyramidal system. This texture coder aims at visual quality enhancement at medium/high bit-rates. Therefore, the method offers a natural basic SNR scalability.

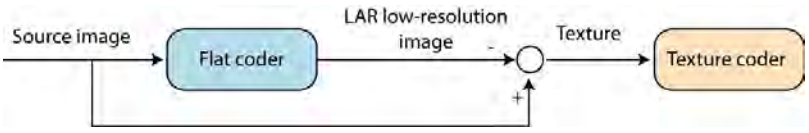


Figure 2.1: General scheme of two-layer LAR coder

The LAR codec tries to combine both efficient compression in a lossy or lossless context and advanced functionalities and services as described before. To provide a codec which is adaptable and flexible in terms of complexity and functionality, different tools have been developed. These tools are then combined in three profiles in order to address such flexibility features (figure 2.2).

Therefore, each profile adresses to different functionalities and different complexities. The Base-

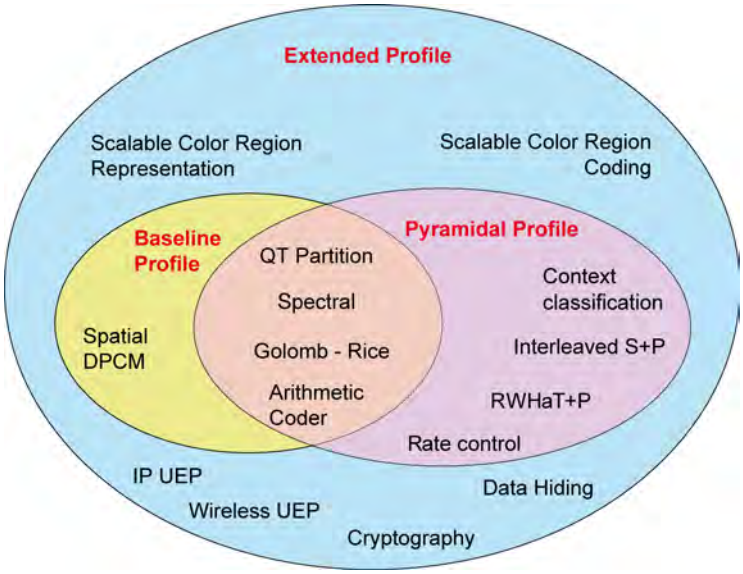


Figure 2.2: Specific coding parts for LAR profiles

line profile is related to low complexity, low functionality requirements, the Pyramidal profile shows an increased complexity but new functionalities such as scalability, rate control are available, and the Extended profile includes scalable color region representation and coding, cryptography, data hiding, unequal error protection services, naturally leading to higher complexity.

An extensive part of my work has been conducted within the medical image compression context. As a consequence, the baseline profile, dedicated to low bit-rate encoding, is clearly not appropriated. As medical image compression requires lossless solutions, we then focus the discussion on functionalities and technical features provided by the pyramidal and extended profiles dedicated to content protection: cryptography, steganography, error resilience, hierarchical securized processes. In this context, the Interleaved S+P coding tool [21] has been chosen as the appropriate tool.

2.2.1 Principles and properties

As previously mentioned, a two-layer framework has been designed. This image decomposition into two sets of data is thus performed conditionally to a specific quadtree data structure, encoded in the Flat coding stage. The basic idea is that local resolution, in other words the pixel size, can depend on local activity. Thanks to this type of block decomposition, block size implicitly gives the nature of the given block: smallest blocks are located upon edges or on texture areas whereas large blocks map homogeneous areas (figure 2.3). Then, the main feature of the Flat coder consists of preserving contours while smoothing homogeneous parts of the image.

This quadtree partition remains the key system of the LAR codec. Consequently, this coding part is required whatever the chosen profile.

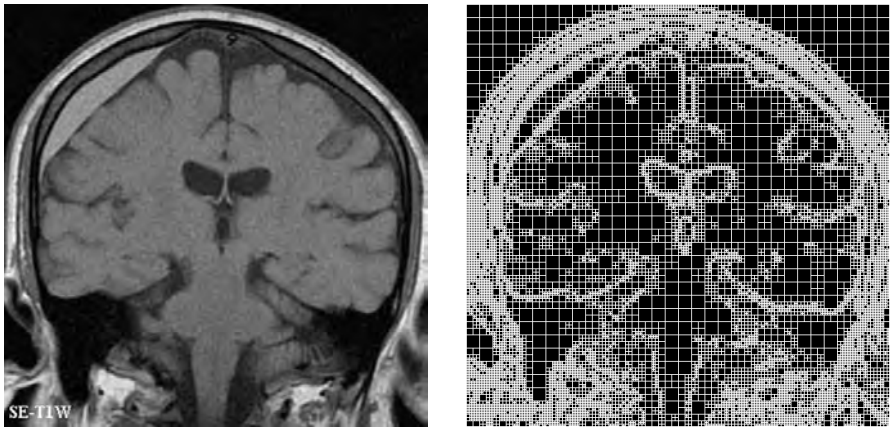


Figure 2.3: Original image and associated quadtree partitions obtained with a given value of activity detection parameter

2.2.2 Lossy to lossless scalable solution

Scalable image decompression is an important feature as soon as very large images are used. Scalability enables progressive image reconstruction by integrating successive compressed sub-streams during the decoding process.

Scalability is generally first characterized by its nature: resolution (multi-size representation) and/or SNR (progressive quality enhancement). The LAR codec supports both of them but differs from JPEG2K. The main difference is that the LAR provides a multiresolution solution together with an "edge oriented" quality enhancement. The lossy or lossless coding process involves two-pass dyadic pyramidal decomposition (figure 2.4). The first pass, leading to a low bit-rate image, encodes the overall information in the image thanks to the Flat coder, thus preserving main contours, while smoothing homogeneous areas. The second pass using a Texture coder adds the local texture in these areas. The LAR framework also contains some interpolation / post-processing steps that can smooth homogeneous areas while retaining sharp edges.

The second important feature for scalability concerns granularity. Scalability granularity defines which elementary amount of data can be independently decoded. Among existing standards, JPEG2K offers the finest granularity of scalability. On the other hand, JPEG provides no scalability, except in its progressive mode, while JPEG-XR enables up to 4 scalability levels. With the LAR codec, the number of dyadic resolution levels N is adjustable, with two quality levels per resolution. Therefore, the number of elementary scalable sub-streams is equal to $2N$.

2.2.3 Hierarchical colour region representation and coding

Within the extended profile, for colour images, we have designed an original hierarchical region-based representation technique adapted to the LAR coding method. An initial solution was proposed in [40] that can be seen as an adaptation of the split/merge methods that tackles coding constraints. .

Even if efficient context-based methods adapted to quadtree-based region partition compression have been developed [138], prohibitive partition coding cost stays one of the principal restrictions to the evolution of content-based coding solutions. To avoid the prohibitive cost of region shape

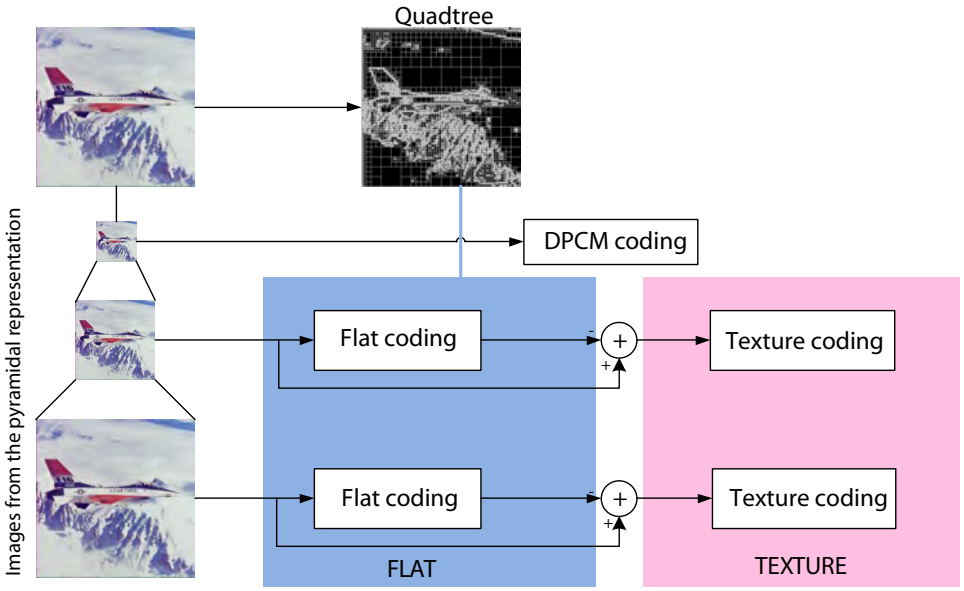


Figure 2.4: Pyramidal representation of an image

descriptions, the most suitable solution consists of performing the segmentation directly, at both the coder and decoder, using only a low bit-rate compressed image resulting from the Flat codiing layer. The segmentation process allows then several simultaneous merges in order to limit complexity and to provide a more compact multi-scale representation. Moreover, a joint mean/gradient criterion weighted by region surfaces (non-symetrical distance) has been defined in order to favor regions with spatial consistency. By iteratively increasing thresholds, a hierarchical segmentation is obtained and allows to efficiently describe the image content from finest to coarse scale. Having as many indexed levels as threshold levels, indexed hierarchical segmentation can be described with a N-ary tree structure called Partition Tree PT_s^N (s: spatial; st: spatio-temporal) where N is the number of indexed levels [122][168]. The multi-scale representation is said to be a self-extracting process (cost free) because both coder and decoder only work on Y-block image.

Color information is used to improve segmentation quality and the encoding of the chromatic components using region representation. To take advantage of color information in the chromatic components encoding, a chromatic control principle is defined and included in the merging process previously described. This chromatic control generates additional binary information for each luminance-based merging attempt to control the merging process. Natural extensions of this particular process have also made it possible to address medium and high quality encoding and the region-level encoding of chromatic images. Another direct application for self-extracting region representation is found in a coding scheme with local enhancement in Regions Of Interest (ROI).

2.2.4 Interoperability

The extended profile also proposes the use of dedicated steganography and cryptography processes, which will be presented in chapter 4. To sum up, the interoperability of coding and representation operations leads to an interactive coding tool. The main features of the LAR coding parts are depicted in figure 2.5.

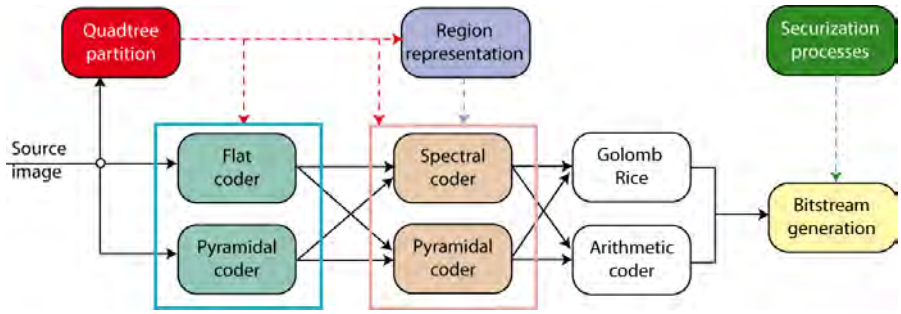


Figure 2.5: Block diagram of extended LAR coder profile

2.3 Quadtree Partitioning: principles

Systems based on a variable-size block representation rely on the knowledge of both a homogeneity criterion and a specific partition. To avoid overlapping, a common partition solution follows quadtree topology.

The LAR framework involves a quadtree partitioning $P^{[N_{max}...N_{min}]}$ with all square blocks having a size equal to a power of two, where N_{max} and N_{min} represent respectively the maximum and minimum authorized block sizes. Thus, the partitioning process consists of first splitting the image into uniform N_{max} square blocks and then building a quadtree on each block.

In fact, many methods rely on such a variable-size block representation. In particular, MPEG4-AVC/H.264 intra mode authorizes a partition $P^{[16,4]}$ (it splits images into 4×4 or 16×16 blocks), where size selection operates to produce the best bit rate/distortion from a PSNR point of view [70]. Methods based on tree structure operate from the highest level (or maximal size) by cutting nodes down into sons when the homogeneity criterion is not met. Although several homogeneity tests can be found in literature [173, 184], in most cases they rely on a computing stage of a L_1 or L_2 norm distance obtained between the block value and the value of its four sons.

In this section, different solutions of designing the mandatory homogeneity criteria are presented.

2.3.1 Basic homogeneity criterion: morphological gradient

A first criterion, based on edge detection, as been especially elaborated for LAR coding purposes. Among the various possible filters, a morphological gradient filter (the difference between maximum and minimum luminance values on a given support) has been chosen, because of its fast, recursive implementation and the resulting limitation of the absolute value of texture.

General description. Let $I(x, y)$ represents a pixel with coordinates (x, y) in an image I of size $N_x \times N_y$. Let $b^N(i, j) \in I$ be the block of size $N \times N$.

Let $P^{[N_{\max}...N_{\min}]}$ be a quadtree partition, and $\min[b^N(i, j)]$ and $\max[b^N(i, j)]$ be respectively the minimum and maximum values in block $b^N(i, j)$.

For each pixel, the block size is given by

$$Siz(x,y) = \begin{cases} \max(N) & \text{if } \exists\, N \in [N_{\max} \dots N_{\min}] : \\ & \left|\max\left[b^N\left(\lfloor\frac{x}{N}\rfloor, \lfloor\frac{y}{N}\rfloor\right)\right] - \min\left[b^N\left(\lfloor\frac{x}{N}\rfloor, \lfloor\frac{y}{N}\rfloor\right)\right]\right| \leq Th \\ N_{\min} & \text{otherwise,} \end{cases} \tag{2.1}$$

where Th represents the homogeneity threshold.

The resulting image of sizes immediately produces a rough segmentation map for the image, where blocks sized with $N_{\min}$ are mainly located on contours and in highly textured areas. This characteristic forms the further basis of various coding and advanced analysis steps provided by the framework.

Color images. For color images, the solution consists of defining a unique regular partition locally controlled by the minimal size among the three components of the considered color space. Then, for each pixel $I(x,y) \in I$, for YCbCr color space, the image of sizes Siz is given by

$$Siz(x,y) = min\left[Siz_Y(x,y), Siz_{Cr}(x,y), Siz_{Cb}(x,y)\right]. \tag{2.2}$$

The thresholds Th can be independently defined for the luminance component and color components. In case of a single threshold Th, the corresponding minimum value (equation 2.2 is mainly supplied from the Y component.

2.3.2 Enhanced color-oriented homogeneity criterion

To address color images representation issue, both a color space and a content extraction strategy has to be defined. As required by JPEG-AIC Call For Proposal, the idea is to add psycho-visual considerations into image codecs together with allowing good rate/distorsion results. Toward this objective, we introduce psycho-visual considerations, especially in the LAR quadtree computation. However, our idea of using the CIELAB color space for analysis can be transposed to other codecs or image processing methods, as soon as they rely in a quadtree representation. This solution has been designed by Clément Strauss during his PhD work in [183, 182].

2.3.2.1 Motivations

Psycho-visual aspects are typically considered for both video coding [10] and image coding such as the JPEG coder [180], where vision models have been used as improvement steps. In particular, works on wavelet quantization noise visibility [200] have led to an perceptual distortion control in JPEG2K [112]. The main drawback of this last approach is that perceptual features are externally used to set up the coder to obtain a given visual quality.

Matching psycho-visual requirements can be then facilitated by means of an appropriate color space. If YUV color space is typically used for video coding purposes, CIE L*a*b* or CIELAB [75][53] is a color space especially designed to exhibit visually uniform color spacing for representation issues. The CIELAB takes into account psychovisual experiments, and models the optical nerve and brain response to color stimuli. Similarly to the Y channel in YUV [204] and YDbDr [66] color spaces, the L* axis of the CIELAB color space represents the lightness, while a* and b* represents chrominances, respectively the red to green axis and blue to yellow axis. The uniform color spacing property of

the CIELAB induces that the Euclidean distance between two colors corresponds to the perceived color difference. This color difference measure has been lately updated by several correcting metrics (ΔE94 [34], ΔE2000 [174]).

Because of its human vision correlation property, we propose to consider the CIELAB color space and to evaluate its capacity to improve the perceptual aspect of the image partitioning when compared to YUV-based representations. In order to increase the perceptual quality, the partitioning process has to give a more accurate representation of the segmentation in particular upon edges. The experiments have demonstrated the representation efficiency of the euclidean distance while keeping LAR compression performances in the same acceptable range.

Let us now define a related homogeneity criterion. All criteria take as inputs the minimum and the maximum pixel values of the three color planes on a given block. For CIELAB color space, $(L_{max}, L_{min}, a_{max}, a_{min}, b_{max}, b_{min})$ values are tested. The homogeneity criterion based on the Euclidean distance is thus defined as $criterion = \sqrt{\triangle L^2 + \triangle a^2 + \triangle b^2}$, where $\triangle L = L_{max} - L_{min}, \triangle a = a_{max} - a_{min}, \triangle b = b_{max} - b_{min}$.

2.3.2.2 Results

Input images are natively in RGB color format and each of these color planes is compressed in the RGB color space by the Interleaved S+P compression algorithm. The image set is composed of five natural images: barba, lena, parrots, peppers, and P06 from JPEG-AIC database [160]. The Interleaved S+P coder is set in a lossless profile and is driven by the quadtree partitioning of the image. This quadtree is computed in either YUV color space or in CIELAB color space. The shown images result from the Flat coder.

Testing procedure. The CIELAB partitioning is set with a threshold T_H giving the closest number of blocks to the number of blocks provided by the ideal YUV partitioning. A reasonable assumption is that an equivalent number of blocks in the CIELAB quadtree space also gives an equivalent quadtree coding cost. This way the representations are compared at the same level of granularity.

Quality metrics. As the LAR framework relies on content-based features, the WPSNR metric must be used with caution [183]. Among perceptual quality metrics modeling subjective quality, the C4 metric has been proved to be the most efficient metric on computing the LAR image quality by correlating the most with human observation [183]. This metric evaluates the quality from 1 to 5 as in subjective assessment rating where 1 represents the poorest quality. This metric is used in the following results.

Comparison of YUV and CIELAB representations. Since bit-rates being the same, the representation quality remains to be evaluated and compared. The representation can be both visually evaluated or by means of the C4 and WPSNR metrics. On some images, the differences are barely noticeable but on others, the CIELAB representation with Euclidean criterion clearly outperforms the YUV representation by introducing less noticeable artifacts.

Visually, the differences usually occur for instance in the red tone color (figure 2.6) and are especially visible in fading areas. YUV partitioning is less discriminant in the red tones and produces

larger blocks while the CIELAB produces smaller and more visually accurate blocks. Figures 2.6 show an example of the superiority of the CIELAB-Euclidean representation.

The visual subjective analysis is correlated with the C4 metric analysis. CIELAB partitioning with the Euclidean criterion produces objectively more visually accurate representation as shown in table 2.1. However, the example images 2.6, while being improved by the CIELAB partitioning, exhibit a lower WPSNR as shown in table 2.2.

C4 measurement	YUV	CIELAB Euclidean
Barba	3.320	3.256
Lena	3.703	3.916
P06	2.953	3.049
Parrots	4.393	4.363
Pimen	2.894	3.038
Average	3.453	3.524

Table 2.1: C4 quality assessments with different partitioning criteria and optimal threshold T_H

Rate/distortion	YUV	CIELAB Euclidean
WPSNR_PIX	24.909dB	24.887dB
Flat bit-rate:	1.243bpp	1.241bpp

Table 2.2: $WPSNR\_PIXRGB$ and bit-rate, YUV and CIELAB representations (image: barba)

Depending on the target application objectives, the CIELAB-oriented criterion can be used. The main drawback is the additional computational cost of the overall process. If low complexity is a required feature as well a low memory storage, morphological based criterion has to be preferred.

2.4 Interleaved S+P: the pyramidal profile

As shown in 2.2, the pyramidal profile relies on the Interleaved S+P framework [21]. In this section, we first introduce the general principles of our algorithm before setting out the construction of the pyramid.

In the following, $M = z_0$ and $G = z_1$ denote the S-transformed coefficients in such a way that if (u_0, u_1) is a couple of value, we have

$$\begin{aligned} z_0 &= \lfloor \frac{u_0 + u_1}{2} \rfloor, \\ z_1 &= u_0 - u_1. \end{aligned} \tag{2.3}$$

2.4.1 General principles

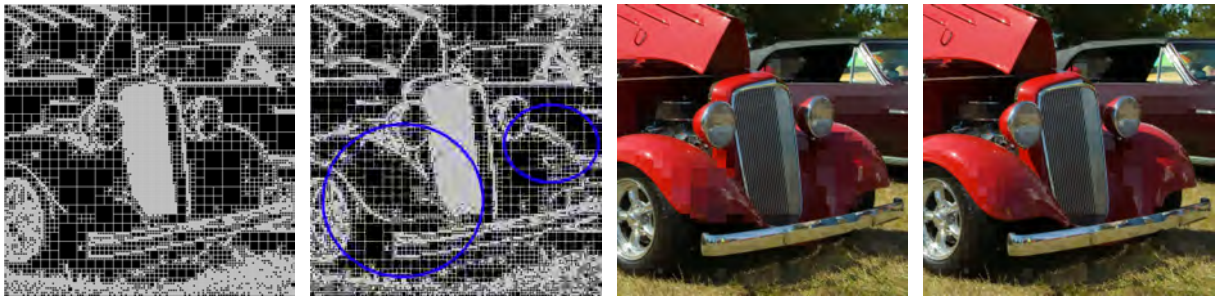
The pyramidal decomposition is based on an adaption of the Wu predictor described in [211]. For full resolution image, errors are coded by means of three interlaced sampling of the original image.

(a) YUV, 29776 blocks

(b) CIELAB, Euclidean criterion, 29746 blocks

(c) YUV block image, 29776 blocks

(d) CIELAB block image, Euclidean criterion, 29746 blocks

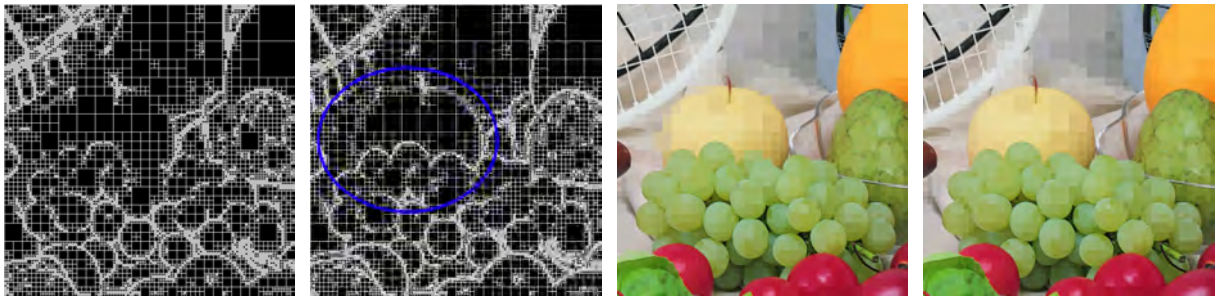


(e) YUV, 20962 blocks

(f) CIELAB, Euclidean criterion, 20932 blocks

(g) YUV block image, 20962 blocks

(h) CIELAB block image, Euclidean criterion, 20932 blocks



(i) YUV, 11779 blocks

(j) CIELAB, Euclidean criterion, 11617 blocks

(k) YUV block image, 11779 blocks

(l) CIELAB block image, Euclidean criterion, 11617 blocks

Figure 2.6: YUV versus CIELAB partitioning.

By this way, we tend to obtain a spatial configuration of 360° type context surrounding a given pixel so that the resulting prediction error is drastically reduced.

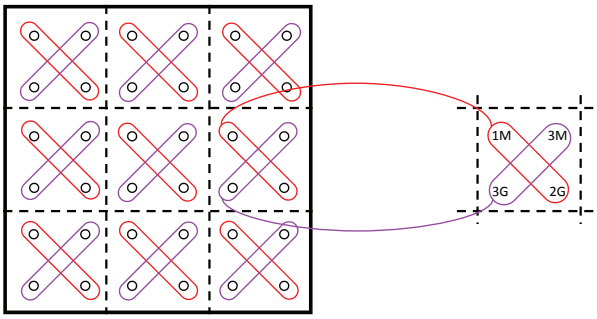


Figure 2.7: Original application of the S-Transform

The general principle of the Interleaved S+P algorithm is as follows. The first step consists of applying the 1D S-Transform on the 2 vectors formed by 2 diagonally adjacent pixels in a 2×2 block, as depicted in figure 2.7. The two vectors of coefficients are sucessively encoded through two successive passes.

2.4.1.1 Notations.

Let define some useful notations.

let $1M$ be the image composed of the z_0 S-transform coefficient of the first diagonal coded through the first pass, such as $1M(i,j) = z(2*i, 2*j)$,
let $2G$ be the image composed of the z_1 S-transform coefficient of the first diagonal coded through the second pass, such as $2G(i,j) = z(2*i+1, 2*j+1)$,
let $3M$ be the image composed of the z_0 S-transform coefficient of the second diagonal coded through the third pass, such as $3M(i,j) = z(2*i+1, 2*j)$,
let $3G$ be the image composed of the z_1 S-transform coefficient of the second diagonal coded through the third pass, such as $3G(i,j) = z(2*i, 2*j+1)$.

2.4.1.2 DPCM principles.

The first pass encodes through a classical DPCM system a uniform subsampled image formed by the average of two diagonally adjacent pixels within each 2x2 block ($1M$ transformed coefficients).

Then the second pass predicts the $2G$ transformed coefficients in order to reconstruct the value of the two pixels of the first diagonal. At this stage, the 360° type prediction context consists of the already known values of the current pass and the diagonal means coded by the first pass.

Finally, the third pass encodes the remaining half of the original image composed of the set of the $3M$ and $3G$ S-coefficients. Once again, thanks to the reconstructed pixels resulting from the two previous passes, a completely spatially enclosing and adjacent context is available to predict the transformed pixel (figure 2.8).

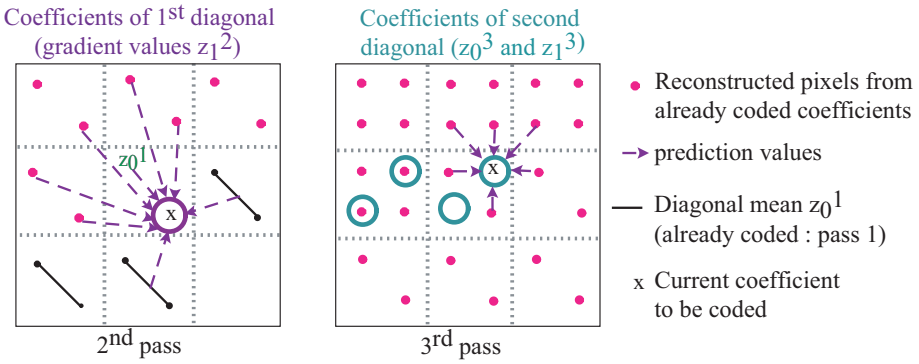


Figure 2.8: Second and third pass of the transformed coefficients prediction process.

2.4.2 Pyramid construction - Interleaving

Let Y the original image be of size $N_x \times N_y$. The multiresolution representation of an image is described by the set $\{I_l\}_{l=0}^{l_{max}}$, where l_{max} is the top of the pyramid and $l = 0$ the full resolution image.

As an extension of the Wu method, four blocks $\frac{N}{2} \times \frac{N}{2}$ are gathered into one block $N \times N$ valued by the average of the two blocks of the first diagonal (first S-pyramid on figure 2.9), that let us write the following equation:

$$\begin{cases} l = 0, & I_0(i,j) = Y(i,j); \\ l > 0, & I_l(i,j) = \left\lfloor \frac{I_{l-1}(2i,2j)+I_{l-1}(2i+1,2j+1)}{2} \right\rfloor, \end{cases} \tag{2.4}$$

with $0 \leq i \leq N_x^l$, $0 \leq j \leq N_y^l$, where $N_x^l = N_x/l$ and $N_y^l = N_y/l$.

The transformation of the second diagonal of a given 2×2 block can also be seen as a second S-pyramid, where the pixel values depend on the ones existing at the lower level of the first S-pyramid. Interleaving is in this way realized.

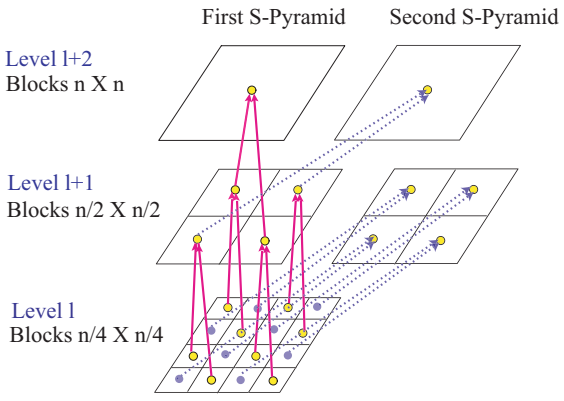


Figure 2.9: Construction of the pyramid

2.4.3 Interleaved S+P Pyramid Decomposition - Refined prediction model

The Interleaved S+P pyramidal decomposition process results from the extension and the adaptation of the Wu prediction method to the whole pyramid. In this section we describe the implemented prediction scheme.

2.4.3.1 Pyramid decomposition principles - Notations

The reconstruction of a 2×2 block located on a given level l is realized in two successive passes. First, the S-transform is applied on pixels $I_l(2i, 2j)$ and $I_l(2i+1, 2j+1)$ so that

$$\begin{aligned} {}^1M_l(i,j) = z_0^{l,1}(2i,2j) &= \left\lfloor \frac{I_l(2i,2j) + I_l(2i+1,2j+1)}{2} \right\rfloor, \\ {}^2G_l(i,j) = z_1^{l,2}(2i+1,2j+1) &= I_l(2i,2j) - I_l(2i+1,2j+1). \end{aligned} \tag{2.5}$$

We immediately can notice that the ${}^1M^l(i,j)$ coefficient is equal to the value of the reconstructed pixel obtained in the upper level of the pyramid leading to

$${}^1M_l(i,j) = I_{l+1}(i,j). \tag{2.6}$$

In a similar way, pixels from the second diagonal of a block are processed as follows:

$$\begin{aligned} {}^3M_l(i,j) = z_0^{l,3}(2i+1,2j) &= \left\lfloor \frac{I_l(2i,2j+1) + I_l(2i+1,2j)}{2} \right\rfloor, \\ {}^3G_l(i,j) = z_1^{l,3}(2i,2j+1) &= I_l(2i+1,2j) - I_l(2i,2j+1). \end{aligned} \tag{2.7}$$

The entire pyramid is then decomposed in two successive descent processes. If the first one is intended to reconstruct the LAR Flat block image, the second one encodes the image texture: the grid information is thus exploited (see figure 2.10).

2.4.3.2 Top of the pyramid - Classical DPCM

On the highest level of the pyramid, we applied the first pass so that the $I_{l_{max}}(i,j) = z_0^{l_{max}-1,1}(2i,2j) = {}^1M_{l_{max}-1}(i,j)$ coefficient values are predicted. For this purpose, we use the classical MED predictor implemented in the compression method LOCO-I [202].

2.4.3.3 LAR block image processing

Reconstructing the LAR block image means that for a given level of the pyramid, a block is processed only if the corresponding block size is lower or equal to the level size (*i.e.* $Siz(x \times l, y \times l) \leq 2^l$). Conversely, if $Siz(x \times l, y \times l) > 2^l$), the block values are copied out so that $I_l(2i,2j) = I_l(2i+1,2j) = I_l(2i,2j+1) = I_l(2i+1,2j+1) = I_{l+1}(i,j)$.

Note that $\breve{z}$ stands for the estimated value of transformed coefficient z.

First S-Pyramid. As the ${}^1M^l(i,j)$ value is already known from the upper level of first S-pyramid, we just have to predict the transformed coefficient ${}^2G^l(i,j)$, standing for gradient value of the first

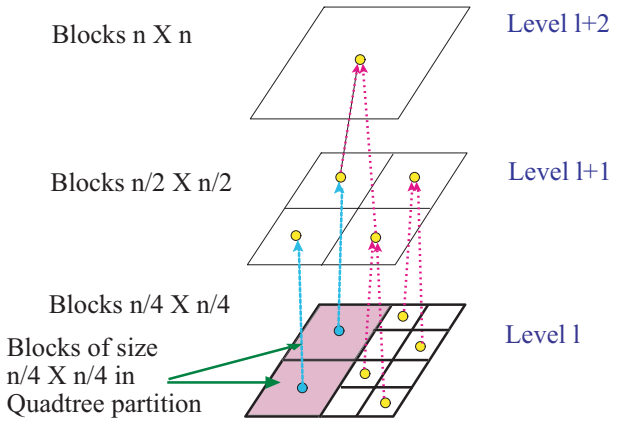


Figure 2.10: Decomposition and extraction of LAR block image data

2×2 block diagonal. In the LAR block image processing context (high activity areas), it turns out that linear prediction is more efficient to estimate $2G^l(i,j)$ leading to

$$\begin{aligned}\widetilde{2G}_l(i,j) = 2.1\Big[0.9I_{l+1}(i,j) + \frac{1}{6}\Big(I_l(2i+1,2j-1) \\ +I_l(2i-1,2j-1) + I_l(2i-1,2j+1)\Big) \\ -0.05\Big(I_l(2i,2j-2) + I_l(2i-2,2j)\Big) \\ -0.15\Big(I_{l+1}(i,j+1) + I_{l+1}(i+1,j)\Big) - I_{l+1}(i,j))\Big].\end{aligned} \tag{2.8}$$

Second S-Pyramid. The third pass of adapted Wu predictor encodes the coefficients from second diagonal. Estimation of $3M^l$ uses inter- and intra-level prediction from reconstructed values, so that

$$\begin{aligned}\widetilde{3M}_l(i,j) = \beta_0^0\frac{1}{4}\Big(I_l(2i-1,2j+1) + I_l(2i,2j+2) \\ +I_l(2i+2,2j) + I_l(2i+1,2j-1)\Big) \\ +\beta_0^1\widehat{1M}_l(i,j),\end{aligned} \tag{2.9}$$

where $(\beta_0^0, \beta_0^1) = (0.25, 0.75)$, and $\widehat{1M}_l(i,j)$ is the reconstructed value of coefficient $1M_l(i,j)$. The estimated value $\widetilde{3G}_l(i,j)$ is computed as follows:

$$\begin{aligned}\widetilde{3G}_l(i,j) = \beta_1^0\Big(I_l(2i-1,2j+1) + I_l(2i,2j+2) \\ -I_l(2i+1,2j-1) - I_l(2i+2,2j)\Big) \\ -\beta_1^1\Big(I_l(2i-1,2j) + I_l(2i-1,2j+2) \\ -I_l(2i,2j-1) - \widetilde{I}_l(2i,2j+1).\end{aligned} \tag{2.10}$$

with $(\beta_1^0, \beta_1^1) = (3/8, 1/8)$. $\widetilde{I}_l(2i,2j+1)$ corresponds to the Wu predictor [211] for the third pass applied to the pixel $I_l(2i,2j+1)$.

2.4.3.4 Texture processing

The second descent process of the pyramid allows to reconstruct the texture. On each level, pixel of size $Siz(x \times l, y \times l) > 2^l$ are here evaluated. Once again, we can distinguish prediction of first S-pyramid from the second one.

First S-Pyramid. Texture blocks are characterized by their low local activity. Thus gradient values $2G(i,j)$ are difficult to estimate through linear prediction. This is the reason why we use median values extracted from close neighborhood.

Let $m_e(u_1, u_2, \ldots, u_n)$ be the median value of a set of n values $(u_1, u_2, \ldots, u_n)$. The estimation of coefficient $2G(i,j)$ located in a texture area is

$$\begin{aligned}\widetilde{2G}(i,j) = \frac{1}{4}\Big(& m_e\big(I_l(2i-2,2j), I_l(2i,2j-2), I_l(2i-1,2j-1)\big) \\ & + m_e\big(I_{l+1}(i+1,j), I_{l+1}(i,j+1), I_{l+1}(i+1,j+1)\big)\Big).\end{aligned} \tag{2.11}$$

Second S-Pyramid. Context is there sufficient to evaluate precisely coefficient values, thanks to linear predictors. $3M_l$ are then estimated through the equation 2.9, where $(\beta_0^0, \beta_0^1) = (0.37, 0.63)$. Predicted coefficients $\widetilde{3G}(i,j)$ are obtained by the application of the relation 2.10, where $(\beta_1^0, \beta_1^1) = (1/4, 0)$.

2.4.4 Quantization Process

The quantization process used in the LAR codec uses quantization factor Q, boundary b_k, reconstruction levels r_k and intervals i_k such as

- $k \geq 0$, $b_k = \lfloor (\frac{Q}{2} + kQ) \rfloor$, $r_k = k$, $i_k = (b_k..b_{k+1}]$,
- $k < 0$, $b_k = \lfloor (\frac{Q}{2} + kQ) \rfloor$, $r_k = k$, $i_k = [b_k..b_{k+1}]$.

Such a quantization process is uniform when Q is odd. However when Q is even the quantization process is uniform with a dead zone around 0. Such a quantization process will naturally have an impact on the statistical properties of the resulting bitstream.

2.5 Joint image coding and embedded systems methodology: an applicative issue

As previously mentioned, the modularity of the scheme authorizes new level of scalability in terms of complexity, closely related to the chosen profile. As long as it is possible, the proposed solutions were developed so that to fit fast prototyping issues. In particular, in connection with the Architecture team of the IETR laboratory, automatic solutions of fast prototyping onto heterogeneous architecture (DSPs, FPGAs), using Algorithm Architecture Matching methodology [67] were proposed. Consequently, if the LAR codec has been developed on PCs, different versions has been implemented onto various embedded systems.

Parallel architectures. First embedded systems oriented works were dedicated to the fast development and implementation of a distributed version of the LAR framework on multi-components platforms [154], or for extended profile with the proper region description, using cosimulation Matlab/C approaches [56]. Solutions for fast design of signal and image processing applications have been then proposed. Within a unified framework, application modeling, cosimulation and fast implementation onto parallel heterogeneous architectures are enabled and help to reduce time-to-market. The overall system provides thus an automatically generated code enabling a high abstraction level for users.

FPGA implementations. In [39], a dedicated FPGA implementation of the Baseline LAR image coder has been designed. In particular, the chosen internal architecture has been designed as a set of parallel and pipelined stages, enabling a full image processing during a unique regular scan. It presents limited requirements for both memory and computation power.The resulting architecture latency remains extremely low as it is determined by the data acquisition for one slice of 8 lines.

Dataflow programming for hardware implementation. Implementing an algorithm onto hardware platforms is generally not an easy task. The algorithm, typically described in a high-level specification language, must be translated to a low-level HDL language. The difference between models of computation (sequential versus fine-grained parallel) limits the efficiency of automatic translation. On the other hand, manual implementation is naturally time-consuming. To solve this problem, designers are establishing solutions to describe the process in a higher level way. In the video coding field, a new high level description language for dataflow applications called RVC-CAL [51] was normalized by the MPEG community through the MPEG-RVC standard [26]. This standard provides a framework to define different codecs by combining communicating blocks developed in RVC-CAL.

In this context, Khaled Jerbi, a PhD student that I partially co-supervised, defined a global design method going from high level description to implementation [88, 89]. The first step consists of the description of an algorithm as a dataflow program with the RVC-CAL language. Then the functional verification of this description using a software framework is realized. The final step consists of an automatic generation of an efficient hardware implementation from the dataflow program. We used this method to quickly prototype and generate hardware implementation of a baseline part of the LAR coder, from an RVC-CAL description. We demonstrated then the ability of the method to provide efficient automatic hardware implementations.

2.6 Conclusion

As described in this chapter, the Interleaved S+P framework tends to associate both compression efficiency with scalable content-based representation. Meanwhile, the JPEG committee has defined the JPEG-AIC guidelines leading to design new or enhance existing functionalities. A large part of my work has been then devoted to JPEG committee through an active participation. Clearly, this time-consuming task could have not been realized without the help of François Pasteau and Clément Strauss (both have defended their dissertation) together with Médéric Blestel, who worked in the IETR laboratory during five years.

Quadtree partitioning remains the key feature of each profile of the LAR method. It acts as

a simplification process of the image, leading to design realistic embedded systems, while keeping interesting compression performances. Nevertheless, JPEG core experiments have shown that the global framework requires advanced improvements in terms of coding features and solution to match easy-to-use application need. These challenges were partially reached through generic coding tools, as described in the next chapter.

Chapter 3

Generic coding tools

Even most part of my work was in connection with the Interleaved S+P coding framework, some advanced solutions were designed so that to be independent of the coding scheme. Indeed, when considering QoS based tools, such as Rate/Distorsion Optimization (RDO) mechanisms, statistical models are necessary. In parallel, from these models, advanced prediction and entropy processes can be imagined. This chapter presents then generic coding tools aiming at enhancing statistical properties as well as pure coding performances.

Among the different functionalities required by JPEG-AIC standardization committee, user-friendly implementation, low computational complexity, as well as efficient multi-component image coding performances, are of first interest. Within joint context of the CAIMAN project and JPEG-AIC standardization process, we designed innovative solutions based of an analysis of the prediction errors of the Interleaved S+P coder. The genericity of the proposed techniques enables to consider them as part of a universal image coding toolbox. Application to the Interleaved S+P codec is however presented throughout this chapter. This chapter corresponds to the main part of François Pasteau's PhD work.

The idea is then to revisit the different coding steps of any predictive coder. First, whereas state of the art codecs usually handle multi component images such as color images using only any color transform, this chapter presents some techniques aiming at improving both the compression ratio and the quality of the compressed color images. These tools exploit the remaining correlation between the components of an image and take advantage on the prediction errors distribution. After defining related notations in section 3.1, adaptive decorrelation has been then designed and is described in section 3.2. This solution relies on improved predictors and is independent of the coding scheme. Therefore they can be applied to any predictive codec such as LAR, H264, etc ...

Secondly, section 3.3 provides a statistical analysis of predictive codecs under the assumption that the prediction error distribution follows a Laplace law. In particular, the quantization process is modeled against the prediction errors to be able to estimate both rate and distortion.

Finally, in section 3.4, the classical entropy coder QM being a binary arithmetic coder, a scan pattern needs to used to encode multiple bit input values. The approach used in the JPEG2K codec, where bit planes are coded one after the other, is not appropriate for the predictive coders which typically require a raster scan. Therefore a symbol oriented QM coding is proposed in this chapter. This scan pattern provides a lower complexity as well as similar compression ratio.

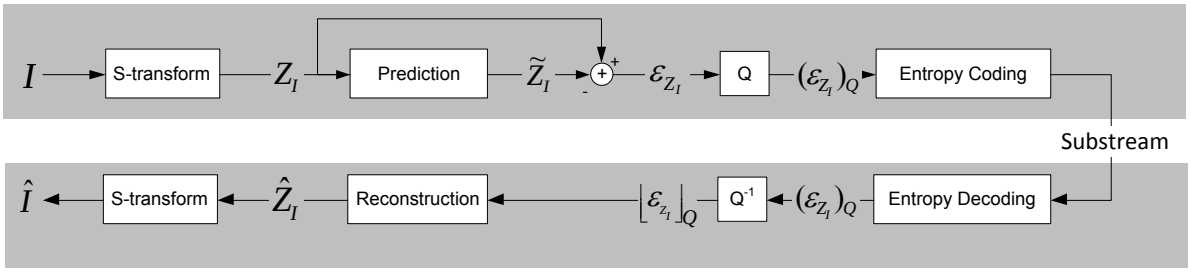


Figure 3.1: Interleaved S+P Coding of a resolution level

3.1 General coding framework - Notations

Figure 3.1 summarizes the coding and decoding processes of an image I through a scalable predictive coder. In this figure, I is an image from a given resolution level of a given component of the original image. It refers to the I_l notation used in the previous section. When considering the Interleaved S+P scheme case, the differences between Flat and Texture coding are not mentioned in this figure as they only change the prediction scheme and the part of the image I_l to encode. This figure introduces the main notations used in this document and shall be referred to.

Let define some notations used in this document.

let I be an image corresponding to a given level of the multiresolution representation of a component of the original image,
let Z_I be one of the transformed image of I, namely $Z_I \in \{2G, 3M, 3G\}$ in case of Interleaved S+P framework,
let $\widetilde{Z}_I$ be the prediction of Z_I,
let ε_{Z_I} be the prediction error between $\widetilde{Z}_I$ and Z_I,
let $(\varepsilon_{Z_I})_Q$ be the quantized version of ε_{Z_I} with quantization factor Q,
let $[\varepsilon_{Z_I}]_Q$ be the reconstructed version of ε_{Z_I} after quantization with respect to Q,
let $\widehat{Z}_I$ be the reconstructed value of Z_I,
let $\widehat{I}$ be the reconstructed value of I.

3.2 Adaptive color decorrelation

3.2.1 Improving decorrelation process: multi component pixel classification

Different studies have been realized so that to lower the resulting bitrate. In particular, solutions based on classification processes have been designed. Indeed, classification can be used to improve the compression ratio of any codec thanks to the principle of conditional entropy. Let A, B and C be three statistical variables. Conditional entropy follows the relation

$$H(C) \quad \geq \quad H(C|A) + H(C|B),$$

where $A \cap B = \emptyset$, $X \cap A \cup X \cap B = X$, $H(X|A)$ being the entropy of X knowing A and $H(X|B)$ being the entropy of X knowing B. Therefore if we can divide error values into different classes thanks to new information given by the context, it would lower the overall entropy and improve the

compression ratio. Classification requires then both context modeling tools as well as classification criteria.

Pixel spatial classification has been first defined in [143] and then is extended to multi-component images through a local activity based classification process [144][145], in the case of the Interleaved S+P framework.

In this document, we focus on the generic multi-component image -based solution.

3.2.1.1 Inter-component classification

As shown in [144], classification performed on pixels according to their neighborhood can improve the compression ratio. The same principle has been extended to multi component images, as for example color images. Such an algorithm tries to use the residual correlation between different components of an image to perform an accurate classification. Actually contours are more prone to have high error values whereas homogeneous areas produce low error values. Those areas tend to be located in the same position between the different image components, leading to propose a dedicated inter-component classification.

Inter-component classification is performed after the prediction scheme as shown on figure 3.2. In this classification scheme, the goal is to classify $(\varepsilon_{Z_C})_Q$ with information coming from $(\varepsilon_{Z_Y})_Q$.

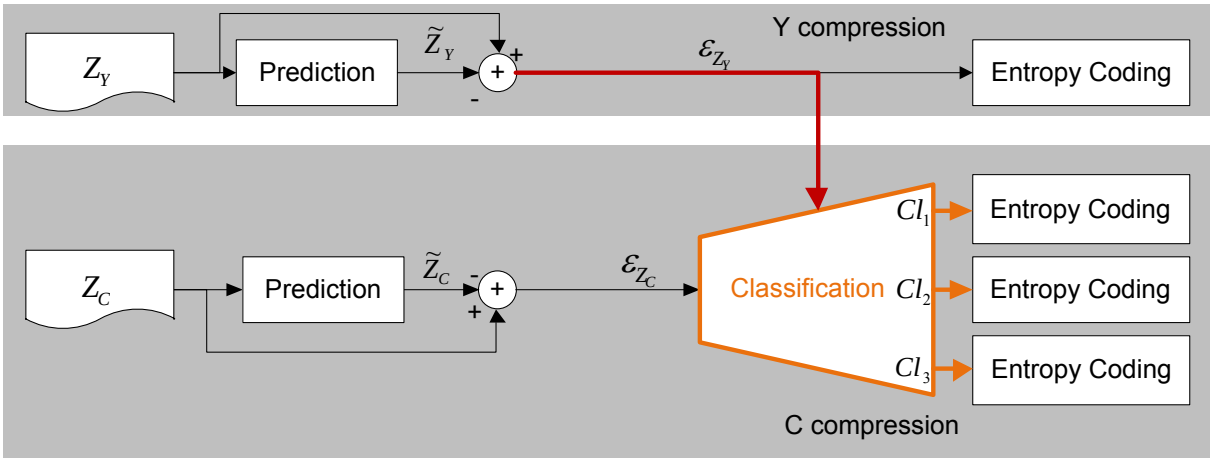


Figure 3.2: Multi component classification scheme

The proposed inter-component classification is performed according to the following algorithm. Let Cl_1, Cl_2 and Cl_3 be the three classes of the classification.

1. One of the three components is used as a reference Y to code another component C, Y coding process is thus not modified.
2. Histogram of absolute values of $(\varepsilon_{Z_Y})_Q$ is computed.
3. A first error threshold $T1$ is set to the histogram mean.
4. A second error threshold $T2$ is set such as $\frac{3}{4}$ of the prediction errors are lower than T2.

5. Classification is then performed as follows:
6. $Cl_1 = \{(\varepsilon_{Z_C})_Q — abs((\varepsilon_{Z_Y})_Q) < T1\}$.
7. $Cl_2 = \{(\varepsilon_{Z_C})_Q — T1 \leq abs((\varepsilon_{Z_Y})_Q) < T2\}$.
8. $Cl_3 = \{(\varepsilon_{Z_C})_Q — T2 \leq abs((\varepsilon_{Z_Y})_Q)\}$.

When considering the Interleaved S+P framework, as the $2G$, $3M$ and $3G$ images present different characteristics in terms of error distribution, the inter component classification scheme is performed independently for each type of errors. It leads to three different classifications for both Flat and Texture coefficients. The aim of such a context modeling is to perform discriminations inside each type of errors. Thus, coefficients of a given type producing the same error distribution are gathered into one context class.

A set of three pictures has been used to evaluate the performances of the proposed solution. This set includes peppers, mandrill and lena pictures (table 3.1). Each of them has different characteristics: mandrill image has high frequencies whereas peppers image features large areas of identical colour. Finally lena has both high frequencies and low frequencies respectively located in her hair and on the background. To obtain comparable results with other image coders such as JPEG2K, a simple arithmetic coder has been used. Four different color spaces were used for the compression, RGB, YCgCo-R [119], O1O2O3 [43] and the Reversible color transform used in JPEG2K [80]. All these color spaces are reversible and can therefore be used for lossless coding [144].

Except for peppers picture, performing colour space decorrelation improves compression results. The YCoCg-R, O1O2O3 and YDbDR colour spaces offer an average 4% gain on mandrill picture and an average 2.5% gain on lena image. The specificity of peppers picture in regard of colour spaces is due to the fact that this image has mainly two predominant colours, red and green. Therefore RGB colour space can, in this particular case, leads to better results. However, even if it produces the best improvement, compression using the RGB colour space remains less efficient in terms of compression ratio than the other colour spaces.

Despite the general use of color images, the development of image codecs such as JPEG, JPEG2K or the newly standardized JPEGXR has been primarily focused on giving the best performance on single component images. To handle color images, state of the art codecs usually rely on color transforms such as YUV, YCgCo and/or subsampling steps to achieve both good compression ratio and good visual quality. However after performing color transforms and/or subsampling, each color component is independently encoded.

When considering lossless coding, coding techniques rely on statistical analysis of the image to perform compression. As subsampling would cause losses, only reversible color transforms can be used. However, even after applying static color transforms, residual correlation still exists between components [144]. This underlaying correlation results in an suboptimal compression rate as decorrelation has been statically done without consideration of the statistics of the image itself. A critical application of lossless coding of colour images concerns cultural digital libraries [116]. Museums actually try to safely digitalize their belongings and thus to produce large quantities of lossless colour pictures. Moreover, current digital cameras are widespread and generate high resolution colour images. Professional photographers tend to prefer lossless compression of their pictures to avoid artifacts due to image compression.

To improve the compression ratio of color images for lossless coding as well as quality for lossy coding purposes, different approaches have been studied. In particular, performing adaptive inter

peppers	Alone	Classification
YCoCg-R	14.975	14.898
O1O2O3	14.968	14.896
YDbDr	15.048	14.960
RGB	14.907	14.767

mandrill	Alone	Classification
YCoCg-R	18.170	18.121
O1O2O3	18.106	18.065
YDbDr	18.122	18.084
RGB	18.948	18.705

lena	Alone	Classification
YCoCg-R	13.582	13.437
O1O2O3	13.536	13.392
YDbDr	13.577	13.422
RGB	13.914	13.662

Table 3.1: Compression results in bpp

component decorrelation was considered. Such an adaptive decorrelation differs from usual reverse color transforms by its automatical adaptivity to the image statistics.

3.2.2 Principles

Adaptive color decorrelation can be seen as a reversible color transform with

$$\begin{pmatrix} Y' \\ C'_1 \\ C'_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -\mu_{C_1} & 1 & 0 \\ -\mu_{C_2} & 0 & 1 \end{pmatrix} \begin{pmatrix} Y \\ C_1 \\ C_2 \end{pmatrix} \tag{3.1}$$

where μ_{C_1} and μ_{C_2} values are adaptively computed during the coding process to best fit the statistics of the image.

The idea is here to take advantage of these statistics so that to improve coding performances. By performing adaptive decorrelation during the prediction process itself, both compression ratio and quality can be improved in a single pass. The figure 3.3 represents the functional implementation of such a technique, whereas the algorithm used to perform the decorrelation is presented in Algorithm 3.1.

On line 11 of algorithm 3.1, the adaptive decorrelation is performed on the predictor applied on component C. The predictor is decorrelated according to the reconstructed value of the prediction error of component Y. The μ_C update process described on lines 15-27 is used to update the

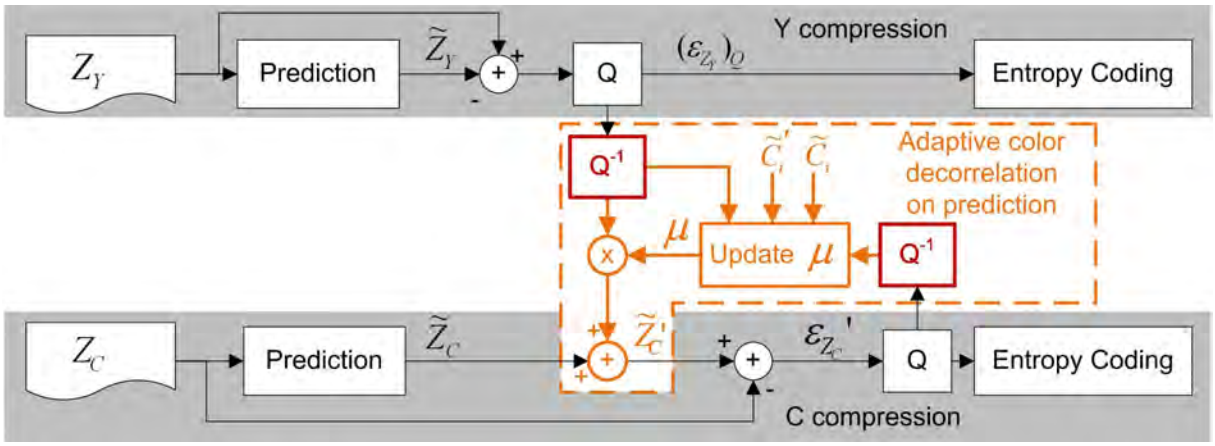


Figure 3.3: Adaptive Decorrelation during Prediction Scheme

decorrelation factor at each iteration by computing a new barycenter of the set of points of coordinates $\{(\widehat{Z_C}' - \widetilde{Z_C}; [\varepsilon_{Z_Y}]_Q)\}$.

To enable a correct behavior at decoder side, $\widehat{Z_C}'$ has to be used as it is the only available reconstructed value at decoder side. However predictions computed before decorrelation $\widetilde{Z_C}$ are used instead of predictions after decorrelation $\widetilde{Z_C}'$ to ensure an accurate update of μ_C.

To avoid $D_C = 0$ we perform a point reflection on the points of coordinate $\{(x; y)|y < 0\}$ around the origin $(0, 0)$. Therefore only the absolute value of $[\varepsilon_{Z_Y}]_Q$ is taken into consideration in D_C. To ensure a locally adaptive decorrelation, we decrease the numerator and denominator by a factor of μ_C after cpt iterations. The number of iterations and the divider of the numerator and denominator have been empirically evaluated.

As adaptive decorrelation can be realized at both encoder and decoder sides, the process itself is costfree. Moreover, due to this symmetric process, adaptive decorrelation is also a lossless process.

3.2.3 Process validation: application to Interleaved S+P codec

Interleaved S+P version [21] of the LAR codec has been used as reference codec to implement and validate this adaptive decorrelation solution. As we showed in a previous study [144] using the RGB color space with an adaptive decorrelation and classification leads to better results than using static color transform. The RGB color space has thus been used during the benchmark process. The image set is composed of 10 images coming from the second core experiment of JPEG-AIC (bike_crop, cafe_crop, p06_crop, p10_crop, rokounji_crop, p26_crop, zoo_crop, green_crop, northcoast_crop, woman_crop)[160]. The resolution of these images is 2 megapixels in 24 bit color depth. Compressions were performed with different quantization factors from lossless Q=1 up to Q=32. To implement inter-prediction, basic predictors directly using the values of the neighborhood were used. Table 3.3 presents average quality results using the WPSNR_MSE metric [177] in dB. Despite of not being directly correlated with the Mean Opinion Score in the case of the LAR codec [183], the WPSNR_MSE metric has been used in these tests as it is one of the quality metrics used in the JPEG AIC standardization process.

Table 3.4 presents average rate results in bits per point (bpp). Column named *Alone* references to

Algorithm 3.1: Adaptive decorrelation algorithm during prediction

{Initialization}
$cpt = 0$
$N_C = 0$
$D_C = 0$
$\mu_C = 0$
for all i **do**
  {Adaptive Decorrelation on predictor}
  $\widetilde{Z_C}' = \widetilde{Z_C} + \mu_C \times [\varepsilon_{Z_Y}]_Q$
  $\varepsilon'_{Z_C} = \widetilde{Z_C}' - Z_C$
  $\widehat{Z_C}' = \widetilde{Z_C}' + [\varepsilon'_{Z_C}]_Q$

  {μ_C update}
  if $D_C > 0$ **or** $[\varepsilon_{Z_Y}]_Q \neq 0$ **then**
    $N_C = N_C + (\widetilde{Z_C}' - \widetilde{Z_C}) \times sign([\varepsilon_{Z_Y}]_Q)$
    $D_C = D_C + |[\varepsilon_{Z_Y}]_Q|$
    $\mu_C = N_C / D_C$
    $cpt = cpt + 1$

    {Ensure local adaptation}
    if $cpt > 1000$ **then**
      $N_C = N_C/4$
      $D_C = D_C/4$
      $cpt = 0$
    end if
  end if
  $entropycoding((\varepsilon_{Z_Y})_Q)$
  $entropycoding((\varepsilon'_{Z_C})_Q)$
end for

the Interleaved S+P codec itself, *Decorrelation during prediction* references to the method presented in previous section.

3.2.3.1 Lossless compression case.

In a lossless context (Q=1), adaptive decorrelation prediction process has been studied with the same experimental conditions as for inter-component classification in section 3.2.1.1. Table 3.2 shows that adaptive inter-component decorrelation improves the compression ratio for all images and colour spaces. The most significant improvements are realized with RGB colour space. RGB being the most correlated colour space used in this work, adaptive inter-component decorrelation can be more accurate and can result in even better compression gain. Inter-component classification leads also to better compression ratios for all images and all colour spaces. It even outperforms the adaptive decorrelation in few cases such as peppers picture with YCgCo-R and O1O2O3 colour spaces. However, in most cases, this classification is still less efficient than the adaptive inter-component

peppers	Alone	Decorrelation	Classification	Dec.+Classif.
YCoCg-R	14.975	14.911	14.898	14.838
O1O2O3	14.968	14.919	14.896	14.854
YDbDr	15.048	14.938	14.960	14.866
RGB	14.907	14.673	14.767	**14.598**
			JPEG 2K	14.920

mandrill	Alone	Decorrelation	Classification	Dec.+Classif.
YCoCg-R	18.170	18.090	18.121	18.053
O1O2O3	18.106	18.067	18.065	18.023
YDbDr	18.122	18.033	18.084	17.986
RGB	18.948	18.012	18.705	17.974
			JPEG 2K	18.176

lena	Alone	Decorrelation	Classification	Dec.+Classif.
YCoCg-R	13.582	13.359	13.437	13.269
O1O2O3	13.536	13.368	13.392	13.265
YDbDr	13.577	13.305	13.422	13.218
RGB	13.914	13.291	13.662	13.207
			JPEG 2K	13.669

Table 3.2: Compression results in bpp

decorrelation, especially on the RGB colour space.

Inter-component classification can be used in addition with adaptive inter-component decorrelation. For all images and colour spaces, compression results are improved when these two schemes are jointly used in comparison with the adaptive decorrelation and the classification used independently.

The best compression scheme in terms of compression ratio is obtained with RGB colour space and both adaptive inter-component decorrelation and inter-component classification. However to obtain better scalability features, YDbDr colour space might be preferred due to its ability to directly retrieve the luminance component.

When compared to the state of the art compression scheme JPEG2K, compression results obtained with the Interleaved S+P scheme, decorrelation and classification are better with each colour space. When using the RGB colour space, 0.32, 0.20 and 0.46 bpp gain can be observed respectively on peppers, mandrill and lena.

3.2.3.2 Lossy compression.

Adaptive decorrelation during prediction process is performed on dequantized prediction errors $[\varepsilon_Z]_Q$ where adaptive decorrelation after prediction process is performed on quantized prediction errors

$(\varepsilon_Z)_Q$. Without quantization, the dequantized value of the prediction error $[\varepsilon_Z]_Q$ is equal to the quantized one $(\varepsilon_Z)_Q$, thus leading to the same results.

However when considering lossy compression, adaptive decorrelation during prediction produces better results and achieves a gain up to 0.4 bpp (20%) for high quantization ($Q = 32$). Adaptive decorrelation during prediction improves both quality and compression rate in all cases.

Quantization	Alone	Decorrelation during prediction
Q = 4	42.709	42.813
Q = 8	38.289	38.494
Q = 32	30.148	30.715

Table 3.3: Quality results of proposed methods in PSNR (db)

Quantization	Alone	Decorrelation during prediction
Q = 1	13.599	12.092
Q = 4	6.6723	5.4472
Q = 8	4.4755	3.5204
Q = 32	1.5075	1.1070

Table 3.4: Rate results of proposed methods in bits per pixel (bpp)

3.2.4 General use of the method

The color-based method of decorrelation defined in this section is entirely reversible. It requires no side information and automatically matches the image content. In terms of performances, the adaptive decorrelation during prediction process decreases significantly the final rate as well as increases the overall objective quality.

Of course, these results have been mainly used together with the Interleaved S+P codec. However, as soon as a predictive codec is defined, the adaptive color decorrelation solution can be integrated within the coding scheme. Typically, the image codec standard JPEG-LS is based on a predictive tool able to discriminate between contours and flat areas: the decorrelation process would provide an enhancement in terms of compression ratios. As for video coders, the MPEG-based framework embeds spatial prediction functional block that could also benefit from our solution.

3.3 Statistical analysis of predictive coders based on Laplacian distributions

In addition to advanced decorrelation solutions, as shown in previous section, complementary coding tools based on statistical modeling process can be designed. In particular, an analysis of predictive coders in terms of error distribution can be conducted RDO purposes.

The proposed analysis relies on a Laplacian probability distribution model that is used to characterize prediction errors according to a reduced set of parameters. From these parameters, estimators of both quality and entropy are proposed. The case of quantized residues is especially considered as quantization process is directly implies for lossy compression purposes. These discrete models are then validated in the case of the Interleaved S+P coder. In this document, only main properties are given: relative proofs are described in [142].

3.3.1 Laplace distribution: a widespread model

When considering image coders, many studies have been done to characterize the distributions of the resulting data. Indeed, whatever the coding principle is (transform-based, predictive or hybrid coder), the idea is to minimize the overall entropy. To this aim, statistical properties are of major interest. Moreover, as we try to achieve realistic solutions of RDO, the chosen statistical model has to be low complex.

In the literature, most of image coder standards are based on Discrete Cosine Transform (DCT). In this context, statistic features of transform residuals have to be estimated [105].

If Gaussian distributions have first been studied, Laplace distributions [97], Generalized Gaussian distributions (GGD) [216][219] and Cauchy distributions [4] lead to more accurate models. As Laplace distribution can be seen as a special case of GGD, the latter shows a higher accuracy in offline analysis [105]. However, tuning GDD requires an additional parameter when compared to Laplace and Cauchy distributions. In [45], authors claim that the predictability of the model parameters is even more important than the accuracy of the models themselves. As a consequence, the high complexity of GDD based analysis prevents from using it for inline RDO solutions even if there is a loss of accuracy. Similarly, even if the generalized Gaussian model gives the most accurate representation of the AC coefficient distribution, the Laplacian model is commonly employed because of its ability to be both mathematically and computationally more solvable [135].

When considering Cauchy distributions, a higher accuracy is observed in the case of heavy tails of transformed residuals [4]. Nevertheless, although Cauchy distribution's mean can be taken as zero, since it is symmetrical about zero, the expectation, variance, higher moments, and moment generating function do not exist. Attempts to estimate these parameters will not be successful. In RDO context, this properties penalize the process and prevent from utilizing it. Laplacian Mixture Model (LMM) have also been proposed for modeling coding residues [141], but the overall performances in both terms RDO and complexity tend to set aside those solutions. When considering wavelet-based coding solution, such as SPIHT-like coders, if hyperbolic probability density function better fits the real distribution [185], Laplacian models are used [199].

As for pure predictive coders, they typically also rely on Laplacian-like error distributions [104][212]. Moreover, classical RDO processes rely on Mean Square Error (MSE) metrics. Even if this metric from maximum likelihood perspectives is not as accurate as the Cauchy metric is [172],

the use of MSE is in particular justified as soon as the error follows a Gaussian or Laplace distribution. To sum up, in the literature, authors usually use Laplace distribution models for coding purposes [212]. A good tradeoff between the computational complexity and accuracy is thus reached.

3.3.2 A Laplace's Law distribution

The first part of this study addresses the problem of finding a probability distribution model which fits the experimental probability distribution of ε_Z. In this first part of the study, quantization is not taken into account to simplify the model (Q=1). Quantization issue will be studied in section 3.3.3.

3.3.2.1 Notations.

We introduce some notations related to the Laplacian probability model.

Let $P_Q(x)$ be the probability $P(x = (\varepsilon_Z)_Q)$ depending on quantization parameter Q,
let α_Q be the scalar such as $P_Q(x) = \alpha_Q e^{-\frac{|Qx|}{b}}$,
let $\widetilde{\alpha_Q}^c$ be the approximated version of α_Q in continuous domain,
let $\widetilde{\alpha_Q}^d$ be the approximated version of α_Q in discrete domain.

3.3.2.2 Laplace probability distribution model

Laplace's law can be expressed by

$$P_1(x) = P\left(x = (\epsilon_Z)_1\right), = \alpha_1 e^{-\frac{|x|}{b}} \tag{3.2}$$

where $P_1(x)$ represents the probability of the prediction error x appearing among $(\epsilon_Z)_1$.

Laplace's law is expressed in the continuous domain. Therefore to ensure fitting to a probability distribution,

$$\int_{-\infty}^{+\infty} \widetilde{\alpha_1}^c e^{-\frac{|x|}{b}} \, \mathrm{d}x \quad = \quad 1 \tag{3.3}$$

should be verified, where $\widetilde{\alpha_1}^c$ corresponds to the approximation of α_1.

Therefore,

$$\begin{aligned} \frac{1}{\widetilde{\alpha_1}^c} &= \int_{-\infty}^{+\infty} e^{-\frac{|x|}{b}} \, \mathrm{d}x \\ \widetilde{\alpha_1}^c &= \frac{1}{2b}. \end{aligned}$$

In continuous domain, a probability distribution model using Laplace's law is approximated by

$$\boxed{P_1(x) \approx \frac{1}{2b} e^{-\frac{|x|}{b}}.} \tag{3.4}$$

3.3.2.3 Discretization of a continuous function issue

Laplace's law is an expression in continuous domain. However, as the probability distribution is computed from discrete values in image compression context, such a function has to be sampled to be a valid probability distribution model. In the rest of the study, N corresponds to the dynamic of prediction errors. Therefore $-\lceil\frac{N-1}{2}\rceil$ and $\lfloor\frac{N-1}{2}\rfloor$ respectively correspond to the lower and higher limit values of these residues.

To discretize the continuous expression between $-\lceil\frac{N-1}{2}\rceil$ and $\lfloor\frac{N-1}{2}\rfloor$, α_1 needs to be reevaluated using the sum $\sum_{x=-\lceil\frac{N-1}{2}\rceil}^{\lfloor\frac{N-1}{2}\rfloor}$ instead of the integral in equation 3.3, so that

$$\sum_{x=-\lceil\frac{N-1}{2}\rceil}^{\lfloor\frac{N-1}{2}\rfloor} \alpha_1 e^{-\frac{|x|}{b}} = 1,$$

, thus leading to

$$\alpha_1 = \frac{1-e^{-\frac{1}{b}}}{1+e^{-\frac{1}{b}}}.$$

Therefore, a probability distribution model from Laplace's law applied to discrete values can be expressed as

$$\boxed{P_1(x) = \frac{1-e^{-\frac{1}{b}}}{1+e^{-\frac{1}{b}}} e^{-\frac{|x|}{b}}.} \tag{3.5}$$

Now that as the expression of the probability distribution model is known, it is important to find the parameters that will ensure a well fitted probability distribution model.

3.3.2.4 Parameters determination

As shown on equation 3.5, Laplace's law probability distribution model can be completely determined by only one parameter b. Therefore, this parameter needs to be chosen to best fit the practical distribution model. To do so, we propose to first determine the α_1 parameter introduced in equation 3.2.

How to determine the α_1 parameter. By noticing that $P_1(0) = \alpha_1$, the determination of α_1 becomes trivial when the whole practical probability distribution is available. Therefore, during the prediction process, it is possible to online determine α_1 by monitoring the number of errors equal to zero produced by the prediction process.

However, such an estimation of α_1 is strongly dependent on the mean value of the probability distribution. In fact, a biased experimental distribution will lead to a completely inaccurate probability distribution model. However as explained before, this study has been performed under the assumption that the mean value of the probability distribution equals zero.

How to determine the b parameter. From α_1, b can be estimated using

$$b = \frac{1}{ln\left(\frac{1+\alpha_1}{1-\alpha_1}\right)}. \tag{3.6}$$

3.3.3 Impact of quantization

Quantization represents a second step in the prediction error production. Such a quantization has been explained in section 2.4.4. First, quantization process has been applied to Laplace's law probability distribution model, and the obtained mathematical relationships are presented here. Secondly, quantization impact on prediction efficiency has been studied.

3.3.3.1 On error distribution

Mathematical approach from Laplace's law. The goal of such a study is to understand the impact of quantization on the probability distribution from a mathematical point of view, without considering the impact of quantization on the prediction efficiency.

As the quantization process is not uniform, it can be mathematically expressed by

$$\boxed{P_Q(x) = \alpha_Q(x) e^{-\frac{|Qx|}{b}},} \tag{3.7}$$

with

$$\alpha_Q(x) = \begin{cases} {\alpha_Q}^0 = 1 - 2\frac{e^{-\frac{\lfloor \frac{Q}{2} \rfloor + 1}{b}}}{1 + e^{-\frac{1}{b}}}, & \text{if x = 0;} \\ {\alpha_Q}^* = \frac{e^{\frac{\lfloor \frac{Q-1}{2} \rfloor}{b}} - e^{-\frac{\lfloor \frac{Q}{2} \rfloor + 1}{b}}}{1 + e^{-\frac{1}{b}}}, & \text{otherwise.} \end{cases} \tag{3.8}$$

When comparing equation 3.2 with equation 3.7, $e^{-\frac{|x|}{b}}$ becomes $e^{-\frac{|Qx|}{b}}$. Without considering $\alpha_Q(x)$, by identification, it is possible to observe that b becomes $\frac{b}{Q}$.

However, as the quantization is not uniform around 0 when the quantization factor Q is even, Because of α_Q being dependant of x, after quantization, the probability distribution does not fit with Laplace's law model anymore.

Approximation of quantized distribution. When quantized, prediction errors do not follow Laplace's law anymore. The approximation $\alpha_Q(x) = \alpha_Q(0) = \widetilde{\alpha_Q}^d$ is then proposed. Under this hypothesis, a probability distribution model following Laplace's law can be found so that

$$\boxed{P_Q(x) \approx \widetilde{\alpha_Q}^d . e^{-\frac{|Qx|}{b}},} \tag{3.9}$$

with

$$\widetilde{\alpha_Q}^d = \frac{1 - e^{-\frac{Q}{b}}}{1 + e^{-\frac{Q}{b}}}.$$

Such a model can be compared to equation 3.5 where b is replaced by $\frac{b}{Q}$

3.3.3.2 On prediction efficiency

As predictions are performed on previously encoded pixels, quantization impacts the prediction error distributions as well as the prediction efficiency. In the previous section, prediction efficiency has not been considered and mathematical relationships were only derived from the unquantized probability distribution model. To assess the impact of quantization on predictors, we define b_Q as the b parameter from the probability distribution model of unquantized errors ε_{Z_I} when quantization factor Q is applied.

Tests have been conducted on the images from the core experiment 2 from JPEG AIC [101]. Each image has been compressed using Interleaved S+P version of the LAR codec using different values of quantization parameter ($Q \in [1, 255]$). For each prediction pass, the probability distribution of prediction errors before quantization, namely ε_Z, has been extracted. Using equation 3.6, b_Q parameter has been estimated for ε_Z. This parameter refers to the efficiency of the predictor. A lower efficiency will be observed as the b_Q parameter increases and an higher efficiency as b_Q decreases.

An increase of the b parameters can be observed, representing the loss of efficiency of predictors. However if we consider the approximation of quantized distribution

$$P_Q(x) \approx \frac{1 - e^{-\frac{Q}{b}}}{1 + e^{-\frac{Q}{b}}} e^{-\frac{Qx}{b}},$$

P_Q can be expressed as a function of $\frac{b}{Q}$ so that $P_Q(x) \approx f(\frac{b}{Q})$. In previous sections, the impact of quantization has not been considered, therefore the approximation $b_1 \approx b_Q$ has been used and an error of $b_Q - b_1$ has been neglected. To estimate the error realized on P_Q, the error $\frac{b_Q - b_1}{Q}$ has been studied. When considering quantized probability distributions, the increase of b related to the lower efficiency of predictors has shown to be compensated by the quantization factor. As a consequence, the impact of quantization on prediction efficiency can then be neglected.

3.3.4 Entropy estimation

From a probability distribution model, entropy can be computed. As entropy computation has a straightforward relationship with the probability distribution model, the accuracy of the entropy estimation relies on the accuracy of the probability distribution model. In this section, three different entropy estimators are proposed from the most simple expression to the most complex one. All estimators are then compared to practical results and discussed.

3.3.4.1 Statistical study

From discrete domain. From the expression of the probability distribution model shown in equation 3.9, entropy H can be estimated using

$$H = \sum_{x=-\frac{N}{2}}^{\frac{N}{2}} -P_Q(x).log_2(P_Q(x)), \tag{3.10}$$

$$H \approx -log_2(\widetilde{\alpha_Q}^d) + 2\frac{\widetilde{\alpha_Q}^d.Q.e^{-\frac{Q}{b}}}{ln(2).b.(1-e^{-\frac{Q}{b}})^2}.$$

Therefore,

$$H = -log_2(\widetilde{\alpha_Q}^d) + 2\frac{Q.e^{-\frac{Q}{b}}}{ln(2).b.(1-e^{-\frac{2Q}{b}})}. \tag{3.11}$$

The first member of the expression namely $-log_2(\tilde{\alpha_Q})$ corresponds to the affine relationship between H and Q. The second member

$$2\frac{Q.e^{-\frac{Q}{b}}}{ln(2).b.(1-e^{-\frac{2Q}{b}})}$$

leads to a positive estimation of the entropy even if $Q > b$.

Equation 3.11 can be expressed as $H = f(Q, b)$. We can notice that

$$\forall \gamma \in \mathbb{R}, f(Q, b) = f(\gamma Q, \gamma b).$$

We consider b^0 and b^1, two possible values for the b parameter and Q^0 and Q^1 two possible values of the quantization factor Q. Therefore if we know $f(Q, b^0)$ for all Q, it is possible to extrapolate $f(Q^1, b^1)$ according to

$$f(Q^1, b^1) \approx f(\frac{Q^1.b^0}{b^1}, b^0).$$

3.3.4.2 Comparison with practical results and limits

Figure 3.4 presents entropy estimations from equation 3.11 as well as the entropy experimentally found. Such an entropy has been computed from the experimental probability distribution without taking into account entropy coding. This experiment has been conducted on the images from the core experiment 2 from JPEG AIC [100]. Each image has been compressed with different quantization factors $Q \in [1, 255]$ and entropy from each stream has been extracted. Models have been constructed from these streams and entropy have been estimated using the proposed models. Figure 3.4 presents then an example of results on a stream reflecting the common behavior of both the models and experimental entropies.

As explained before, experimental entropy has two main characteristics. When $Q < b$, relationship between entropy H and $log_2(Q)$ is affine. However when $Q >> b$, a deadening can be observed. Entropy estimation from continuous domain can not follow the experimental entropy and becomes negative when $Q >> b$. Approximation from discrete domain as well as the mathematical expression from Laplace's law produce similar results and tend to underestimate entropy when $Q >> b$. Such a behavior is to be expected as the proposed models do not take into account the statistic noise. At high quantization factor, this statistic noise will cause a higher entropy as it represents extra information, not considered by Laplace's law models, to be encoded.

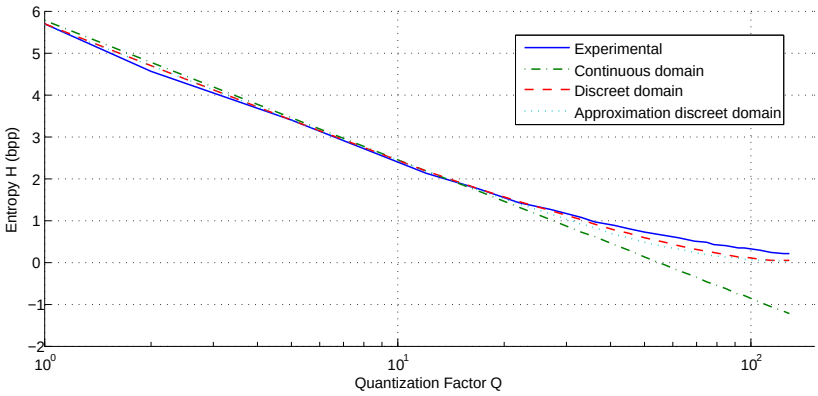


Figure 3.4: Comparison between practical results, pure mathematical, approximated quantized and continuous domain

3.3.5 Mean Square Error estimation

In this section, we estimate the distortion created by the quantization process. To do so, we estimate the mean square error from the probability distribution model and the quantization factor Q.

3.3.5.1 Statistical Approach

Mean square error can be expressed from a Laplace's law probability distribution model according to

$$MSE \approx 4\alpha_1 b^3 + \frac{4\alpha_1 e^{-\frac{\lfloor \frac{Q}{2} \rfloor}{b}}}{1 - e^{-\frac{Q}{b}}} \left(\begin{array}{l} \left(e^{-\frac{1}{b}} - 1\right) b^3 - \left(\lfloor \frac{Q-1}{2} \rfloor e^{-\frac{1}{b}} + \lfloor \frac{Q}{2} \rfloor \right) b^2 \\ \\ + \frac{1}{2} \left(\lfloor \frac{Q-1}{2} \rfloor^2 e^{-\frac{1}{b}} - \lfloor \frac{Q}{2} \rfloor^2 \right) b. \end{array} \right) \tag{3.12}$$

As the MSE estimator has been derived from the probability distribution model without quantization, only $P_1(x)$ has been used and not $P_Q(x)$. Therefore approximations presented in 3.3.3.1 are not used.

All the calculations are performed from the initial hypothesis that the probability distribution follows Laplace's law. Three characteristics can then be observed. First, when quantization factor Q is equal to its maximum value, meaning that all $\widehat{Z_Q} = 0$,

$$\lim_{Q \to \infty} MSE = 4\alpha_1 b^3.$$

Therefore, if $b >> 1$, $4\alpha_1 b^3 \approx 2b^2$, leading to

$$\lim_{Q \to \infty} MSE = 2\alpha_1 b^2.$$

From equation 3.12, MSE can be expressed as a function of Q and b, so that $MSE = g(Q, b)$. Therefore, we can approximate that

$$\forall \gamma \in \mathbb{R} \gamma^2 g(Q, b) \approx g(\gamma Q, \gamma b)$$

As with entropy estimators, if we know $g(Q, b_0)$ for all Q and one value of b_0, it is possible to extrapolate $g(Q_1, b_1)$ using

$$g(Q_1, b_1) \approx \left(\frac{b_1}{b_0}\right)^2 g\left(\frac{Q_1.b_0}{b_1}, b_0\right).$$

Finally, whereas entropy is mean invariant, the efficiency of mean square error estimators relies on the mean value of the probability distribution model being equals to zero.

3.3.5.2 Comparison with practical results and limits

As for entropy estimation, quality estimator has been compared to experimental values. The experiment has been conducted the same way as for entropy estimators. Each image has been compressed with different quantization factors, $Q \in [1, 255]$, and mean square error values from each stream has been extracted. Models have been constructed from these streams and resulting distortions have been estimated using the proposed model.

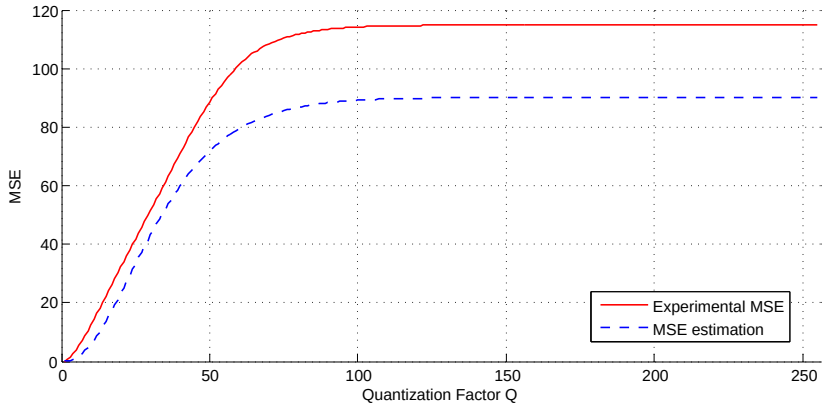


Figure 3.5: Comparison between experimental and estimated MSE

Figure 3.5 presents an example of quality estimation versus experimental quality where limits of estimator can be clearly noticed.

The first limit to be noticed is that the estimation seems to have a bias and a simple translation $Q = Q - 3$ can fix it. This inaccuracy comes from the approximation used to produce equation 3.12, ie $\int$ instead of $\sum$. The maximum distortion is obtained when $Q = 255$. A big difference between the estimated and obtained quality can be nevertheless observed. This difference can be explained by the statistical noise for $Q >> 1$. This statistical noise is mainly dependent on the image and cannot be expressed by the model presented here.

3.3.6 Applications to the Interleaved S+P predictive codec

The described estimators of MSE and entropy for predictive codecs based on Laplacian error distribution are naturally relevant for Rate Distortion Optimization. As JPEG-LS relies on the assumption that residuals follow a two-sided geometric distribution (thus a discrete Laplace distribution) [203], using such estimators would probably take advantage of them espacially for near-lossless purposes.

In [142], an immediate application was the extension of the method to the particular multiresolution description of the Interleaved S+P method. Relations between prediction error substreams of different levels are studied. To this aim, parameters b from substreams of different resolution levels have been compared. Then distortion propagation from one level on another has been studied leading to an estimator of the final mean square error of an image after coding.

From this study, a few mathematical tools have been designed to help realize a good rate distortion optimization. This tools can be divided in three categories: at stream level, at resolution level and from multiresolution. Therefore, we have the following relations.

- at stream level

$$b_{3M} \approx \frac{b_{3G}}{1.7} \tag{3.13}$$

$$H_{Li} = \frac{H_{2G} + H_{3M} + H_{3G}}{3} \tag{3.14}$$

$$MSE_{a\cup b} \approx MSE_M + \frac{MSE_G}{4} + \frac{n_G^i}{4n} \tag{3.15}$$

- for multiresolution

$$b_{Li+1} \approx \frac{b_{Li}}{0.9} \tag{3.16}$$

$$H = \sum_{i=0}^{LMax} \frac{1}{2^i} H_{Li} \tag{3.17}$$

$$MSE_{Li} \approx \frac{1}{2}\left(MSE_{Li+1} + \frac{MSE_{2G}}{4} + \frac{n_{2G}^i}{4n} + MSE_{3M} + \frac{MSE_{3G}}{4} + \frac{n_{3G}^i}{4n}\right) \tag{3.18}$$

Besides rate distortion optimization, these relationships can be used in other domain. For example, the quality of the decoded image can be estimated before the decoding process, only from the bitstream itself. This solution, first developed for JPEG2K in [25], can be then adapted to any coder as soon as estimators of both entropy and quality are known. Such a mechanism estimates the quality of the decoded image without decoding it and without any side information (figure 3.6). First the bitstream is parsed to know the size, the availability of each substream as well as the quantization used. For each substream according to equation 3.6, the b parameter is estimated. From this b parameter and the quantization factor, equation 3.12 is used to estimate the Mean Square Error. Finally equation 3.18 is used to sum the MSE of all substreams included in the input bitstream.

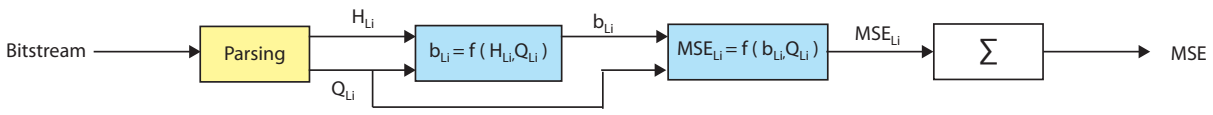


Figure 3.6: A priori quality estimation

3.4 Symbol-oriented QM coding

Sections 3.2 and 3.3 have respectively dealt with the prediction process and the resulting distribution model. The last step of any encoding framework requires an entropy coder to reduce the amount of bytes needed to be stored. It uses a statistical representation of the data to be compressed, so that to produce a bitstream using as little data as possible.

Entropy coding itself can be performed through different entropy coders. Commonly known coders are run length coder, Huffman coder [74], Golomb coder [63] or arithmetic coder [161]. Q15 [81] as well as MQ [80] fall in the last category, i.e. arithmetic coders. Such an arithmetic coder uses the principles of Elias coding using intervals to represent input symbols. The size of each interval corresponds to the probability of the symbol to appear. Q15 and MQ belong to a subcategory of arithmetic coder, namely binary arithmetic coder. Such coders use only two symbols as input, 0 or 1 each represented by a separate interval.

Arithmetic coders generally offer good compression ratio as well as a relatively low complexity in regards to other arithmetic coders. To lower the overall complexity, the idea is here to propose a symbol-oriented coding scheme based on QM coding scheme.

3.4.1 Symbol-oriented entropy coding: motivation

As a binary arithmetic coder can only take binary values (0 or 1) as input, a scan pattern is required to encode values coded with multiple bits. The JPEG2K entropy coder uses a bit-plane oriented coding [169]. Bit plane oriented coding involves coding bits in an "horizontal" way: first all most significant bits from all values are encoded, then the next less significant bit plane until it reaches the least significant bits. As JPEG2K is not a predictive codec and relies on a wavelet transform, an approximated decoding of the values from only the few first bits already gives interesting information to reconstruct the pixel values. Therefore such a bit plane coding can be used as part of a rate control. However in case of predictive codecs, fully reconstructed prediction error values are required. Indeed an approximated prediction error would result in incorrect pixel reconstruction as well as incorrect predictions afterwards, leading to error propagation within the reconstructed image.

Moreover a bit plane oriented coding needs the whole input stream to be available to start encoding. Classically, prediction errors are computed one after another following a raster scan. This way, the bit plane oriented coder has to wait that all predictions have been realized before starting and therefore does not allow parallelism between prediction and entropy coding (figure 3.7).

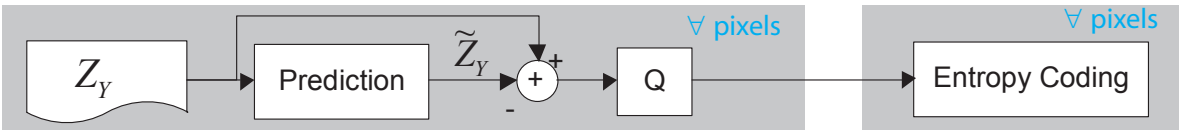


Figure 3.7: QM coding scheme with bit plane orientation

To address this issue, a symbol-oriented QM coding is proposed. Such a symbol oriented coding involves to code bits in a "vertical" way. All bits from an input value, referred as a symbol in the case of entropy coding, are then encoded before starting encoding all bits of the next symbol. Such a scheme enables inline entropy coding of prediction errors directly after the prediction process (figure

3.8). Such a characteristic enables

- to monitor the stream size for rate control and adjust the quantization factor through a feedback loop,
- to determinate the context for the QM Coding during the prediction process,
- to pipeline the prediction process with the entropy coding.

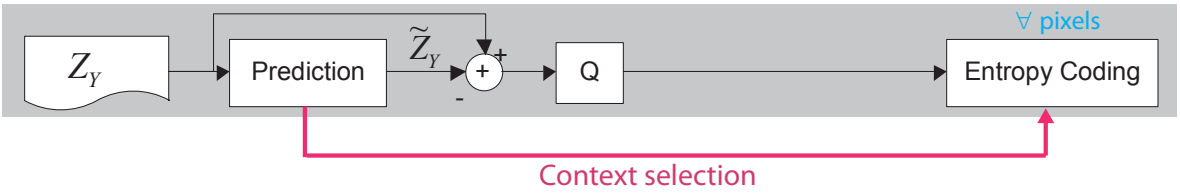


Figure 3.8: QM coding scheme with symbol orientation

3.4.2 QM bit plane oriented coding: an overview

As shown on figure 3.9, QM bit plane oriented coding is performed in an "horizontal" way. To avoid being dependent of the sign, symbols are not coded in a 2's complement representation but rather with a sign bit and the absolute value of the symbol. This entropy coding framework relies on three different passes: significant pass, sign coding and magnitude refinement pass.

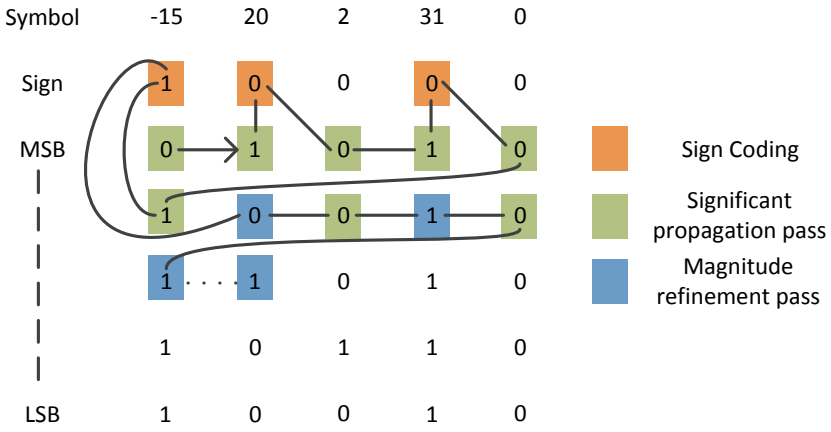


Figure 3.9: Example of QM coding with bit plane orientation

Significant pass. This pass aims at encoding the first most significant bits of the symbols. The bit plane oriented coder starts in significant pass that processes the most significant bit of the symbols. Two cases can arise after coding a bit:

- After coding a "0" value, the coder continues in the same bit plane with the next symbol.

- After coding a "1" value, the coder encodes the sign of the current symbol in the sign coding pass. All remaining bits of this symbol will be from now on coded in the magnitude refinement pass. After coding the sign of the symbol, the coder continue in the same bit plane with the next symbol (symbol '20' in figure 3.9).

Sign coding. As explained previously, sign coding is performed when a "1" is met in the significant pass. Sign coding is performed by encoding the binary value of the sign with the MQ Coder.

Magnitude refinement pass. Finally magnitude refinement pass is performed for all remaining bits after finding a "1" in the significant pass.

3.4.3 Proposed symbol oriented coding

As shown on figure 3.10, symbol plane oriented coding is performed in a "vertical" way. As for bit plane oriented coding, symbol oriented coding uses a sign bit and the absolute value of the symbol. It also involves three different passes: magnitude, sign and refinement coding.

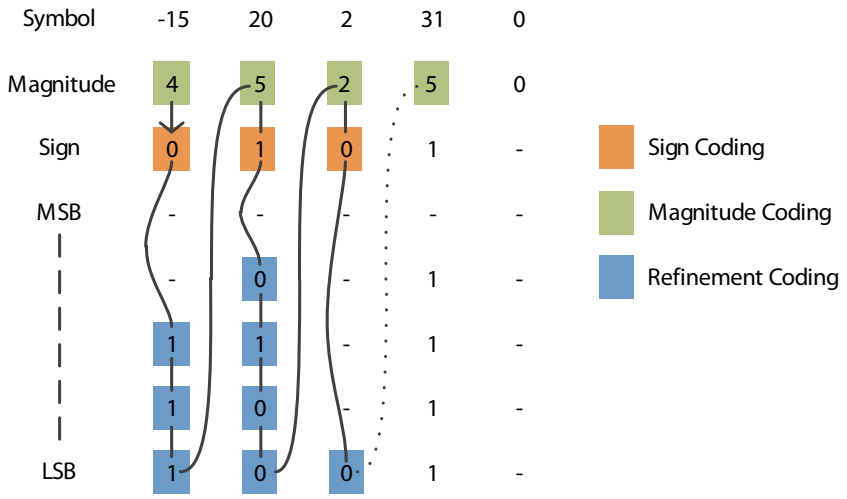


Figure 3.10: Example of QM coding with symbol orientation

Magnitude coding. Magnitude coding (figure 3.11) is designed so that to encode the number of bits needed to represent the symbol to be encoded as well as to adjust the complexity of the overall entropy coding process by minimizing the number of QM coding passes used.

Magnitude coding is performed in three steps.

- First, the minimal number of bits $\overline{X}$ needed to represent the symbol X to encode is computed, as shown in algorithm 3.2. It relies on a loop to look for the minimal number of bits needed to encode the symbol X. Let $\max_{\overline{X}}$ be the maximal possible value of $\overline{X}$. This loop is performed at most $\max_{\overline{X}}$ times and requires only bit shifting and subtraction. Therefore its complexity remains low.

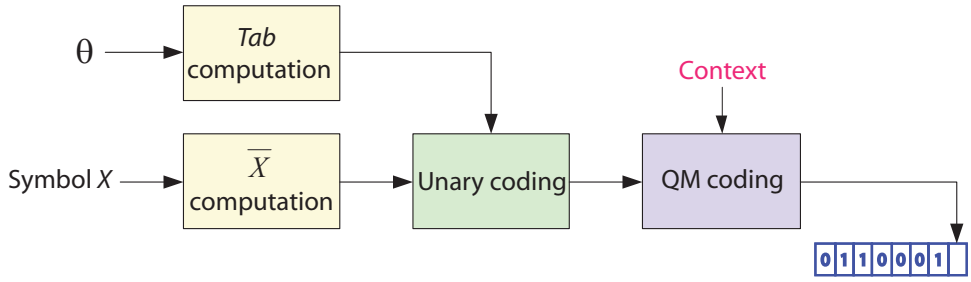


Figure 3.11: Magnitude coding process

- Then $\overline{X}$ is coded through unary coding with the dictionnary Tab containing the length of the codeword. To be able to construct a well fitted dictionary, a statistical study of the symbol to encode (i.e. prediction errors) has to be carried out (typically as described in section 3.3). From this study, the probability distribution of the magnitude of the prediction errors can be estimated. Let $P_{Mag}(x, \theta)$ be the probability $P(x = \overline{X})$ with an error probability distribution estimated with the set of parameters θ. The dictionary Tab can be then determined using algorithm 3.3 as well as the error distribution model $P(x = \overline{X})$.
- Finally, each bit from the unary codeword is encoded with the QM Coder.

3.2.

Algorithm 3.2: $\overline{X}$ Computation

Require: X
Ensure: $\overline{X}$
1: {Initialisation of a temporary variable} $comp = 2^{\max_{\overline{X}}}$
2: $\overline{X} = \max_{\overline{X}}$
3: **while** $X < comp$ **do**
4: $comp = \frac{comp}{2}$
5: $\overline{X} = \overline{X}$ - 1
6: **end while**

This algorithm is relatively complex as it involves sorting. However the sorting mechanism is applied on a reduced table of $\max_{\overline{X}}$ elements. For example if we consider an 8 bits per component image, $\max_{\overline{X}} = 10$, leading to a sorting process on only 10 values. When knowing θ, all possible codeword lengthes in Tab can be determined before actually performing the entropy coding and does not have to be computed for each symbol to encode. However, if the entropy coding is performed inline with the prediction, the θ parameters of the whole stream is naturally unknown at the time of the entropy coding.

Different solutions to this problem can be used. In particular, to respect a low complexity constraint, an estimation of the $\frac{b}{Q}$ from previously encoded stream is first used to estimate Nb parameters. Then after encoding a certain number of symbols, the set θ of parameters are again approximated but this time using the previously encoded prediction errors. This scheme enables a low complexity approach as well as an adaptivity to the stream statistics during coding.

Finally each bit from the unary codeword is encoded using the QM coder according to a given

Algorithm 3.3: Tab Computation

Require: θ, $\max_{\overline{X}}$
Ensure: Tab
1: {Initialization} Tab = [0 0 0 ... 0]
2: **for** $i = 0$ to $\max_{\overline{X}}$ **do**
3: Compute $P_{Mag}(i, \theta)$
4: $j = i - 1$
5: {Sorting $P_{Mag}(i, \theta)$ among the already computed $P_{Mag}(Tab[j], \theta)$}
6: **while** $j \geq 0$ and $P_{Mag}(i, \theta) > P_{Mag}(Tab[j])$ **do**
7: $Tab[j + 1] = Tab[j]$
8: $Tab[j] = i$
9: $j = j - 1$
10: **end while**
11: **end for**

context. Such context modeling technics are explained in the next section.

Sign coding. Sign coding is performed directly after magnitude coding. The algorithm here does not differ from classical bit plane oriented QM Coding. Only the context modeling varies and is presented in the next section.

Refinement coding. The remaining bits are then encoded. The number of refinement bits has been previously determined by the magnitude coding pass.

After having performed these three coding passes on a symbol, the whole process is repeated on the next symbol until the end. Next section will discuss about the context modeling techniques used for each coding pass to improve the efficiency of the entropy coder.

3.4.4 Context modeling

The context modeling techniques used in the symbol oriented QM coder are based on spatial classification of prediction errors. When comparing symbol oriented and bit plane oriented coding schemes, the context orientation is different. If we consider bit plane oriented coding, a $360°$ context is available from higher bit planes. As for the symbol oriented coding, due to the raster scan, only a $180°$ neighborhood is available to construct the context model. However, this context can be improved into a $360°$ context by using values from the directly higher resolution level.

Magnitude coding. The classification process uses a $360°$ context with three thresholds to obtained four different classes. Using four classes was empirically determined to be a tradeoff between a good context modeling and context dilution. As it encode the unary codeword, QM coding process decreases the context number for each bit it encodes. Therefore the context values need to be distant from at least the maximum magnitude $\max_{\overline{X}}$ that can be coded. As an example, in case of 8 bits per component images, there should be at least 10 contexts in each class.

Sign coding. The classification process also uses a 360° context to track the sign of the prediction errors of the neighborhood. From this neighborhood of 8 values, an average value is computed and 5 classes are generated. The first class corresponds to the case when 4 values are positive and 4 values are negative. The fifth class corresponds to the case when the eight values share the same sign. To take into account the most probable sign, the sign of the prediction error to encode is XORed with the most probable sign from the surrounded context.

Refinement coding. Considering the refinement pass, no context modeling has been used as of yet. Therefore there is only one context for this pass.

3.4.5 Validation: application to the Interleaved S+P

For validation purposes, the proposed QM coding scheme has been implemented in the Interleaved S+P codec. Each substream has been coded according to both symbol oriented and bitplane oriented coding schemes. Entropy coding is performed inline with the prediction process.

To compare both complexity and compression ratio, the set of images of the core experiment 2 of JPEG AIC standardization process has been used. For each image, 30 compressions have been performed from low to high bitrates. Both symbol and bit plane oriented coding have been used. Concerning the complexity benchmark, results have been obtained by performing each compression 50 times and taking the average time of these 50 compressions.

Compression ratio. Figure 3.12 presents the average compression ratio of both bit plane oriented and symbol oriented coding. An overhead between 0.5% and 3.5% can be observed for the proposed method against bit plane oriented coding. This result could be mainly improved by using a better context modeling. The context modeling proposed in this chapter is a preliminary study and additional work should be done to improve the results.

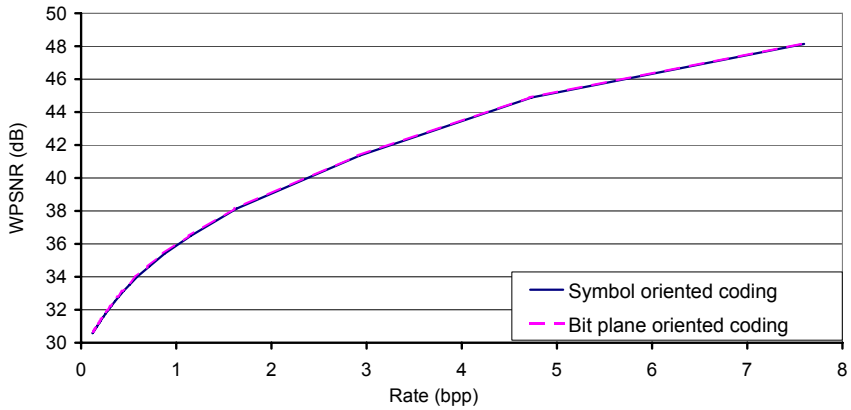


Figure 3.12: Compression ratios of QM encoding for both symbol oriented and bit plane oriented QM Coding

Complexity. Figure 3.13 presents the computational time of both bit plane oriented and symbol oriented coding for all images and three compressions. The average gain in computational time is

around 50% for the entropy coding process and 33% for the whole encoding process of the Interleaved S+P coder for low bitrate as well as for high bitrates.

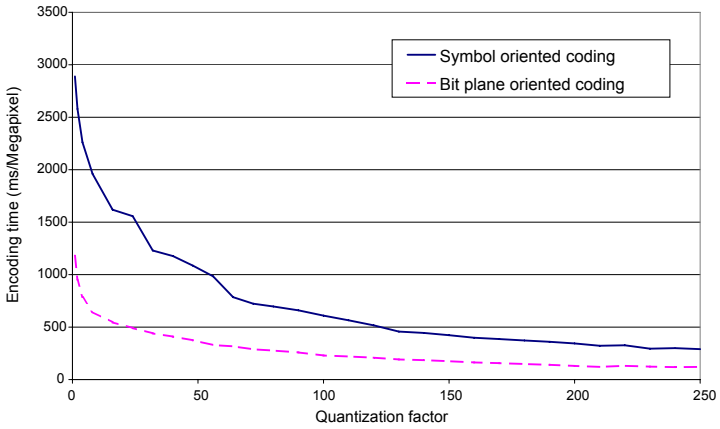


Figure 3.13: Compression ratios of QM encoding for both symbol oriented and bit plane oriented QM Coding

3.5 Conclusion

In this chapter, we presented three different generic tools for coding purposes. First, based on an inter-component adaptive decorrelation, multi-component compression framework has been presented. Results obtained show bit-rate saving from 0.20 up to 0.46 bpp have been observed. Considering colour spaces, using the RGB colour space together with the decorrelation process leads to the best compression results. However, for better scalability features, YDbDr colour space might be considered.

Secondly, a statistical analysis of predictive coders based on Laplacian distributions provides estimators of entropy and distortion. These tools can be then combined for advanced functionalities such as rate control and rate distortion optimization processes.

Finally, a symbol oriented QM coder was then proposed and compared to the classical bit plane oriented QM coder. Results in terms of complexity of the proposed scheme were shown to be better than bit plane oriented coding with a gain of about 60 %.

If the validation process has been conducted with the Interleaved S+P coder, as previously mentioned, the genericity of the presented techniques allows to envisage their integration into other predictive coders.

As we shown in this chapter, this generic coding toolbox should enhance compression performances as well as providing advanced Quality of Service (QoS) features based on statistical modeling of prediction errors. In particular, improving Rate / Distorsion features naturally enhances Quality of Experience. Even if the quality metric used in this study remains the debatable MSE, the related complexity insures realistic network based implementations.

Along with these coding-oriented tools, additional services has to be designed so that to guarantee an end-to-end QoS. In particular, the next chapter focuses on data integrity.

Chapter 4

Content securization and Quality of Service: preserving end-to-end data integrity

Quality of Experience is not only based on the lone received image quality. Typically, medical images or high resolution art images usually embed private metadata that have to remain confidential. Huge amount of medical data are then stored on different media and are exchanged over various networks. If these embedded sensitive data are accidentally altered, even if the received image quality is sufficient, end-users can perceive the overall process as an inefficient one. As a consequence, techniques especially designed for these data are required so that to provide security functionalities such as privacy, integrity, or authentication. Multimedia security is thus aimed towards these technologies and applications [24]. In addition, to insure reliable transfers, flexible and generic scheduling and identification processes have to be integrated for database distribution purposes while taking into account secure remote network access together with future developments in network technologies.

Then, on one hand, content protection consists of preserving data integrity and masking data content. The commonly used methods to obtain these protections are respectively hashing and ciphering. On the other hand, the embedding of hired data aims to protect copyrights or add metadata to a document. Besides watermarking, steganography, and techniques for assessing data integrity and authenticity, providing confidentiality and privacy for visual data is among the most important topics in the area of multimedia security. Applications range from digital rights management to secured personal communications, such as medical materials.

Moreover, errors can occur in various ways when transmitting multimedia contents. Depending on the communication media, i.e. cable network, wireless network (wifi, cellphone network) or even during physical storage (hard disk, flash memory...), binary or packet errors can appear, the worst cases being potentially the wireless network followed by the cable network. In order to provide the best user experience, error concealment and robustness strategies are incorporated during both source coding and the transmission process. As for the channel coding, protection strategies depend on transmission conditions in order to ensure Quality of Service (QoS).

When considering the source coding side, this robustness is usually obtained by adding a given level of redundancy within the transmitted data. Under certain conditions this redundancy allows to recover the whole information even if some parts of the information are damaged. Hybrid approaches,

named joint source-channel coding, achieve robustness by combining protection strategies at both source and channel coding. However when the errors are too strong, or when data packets are definitely lost, some part of the information can be missing or unusable despite the transmission robustness and the added redundancy.

In this chapter, we address these two different services, namely content protection solutions through Interleaved S+P based mechanisms, and Quality of Services tools designed for MPEG-4 SVC or any scalable image coder. This chapter is organized as follows: section 4.1 defined first the securization oriented application contexts within these works have been realized. Then we focus on two aspects of the data integrity. Section 4.2 provides joint data hiding and cryptography processes, while section 4.3 described two specific tools among Quality of Service ones.

4.1 Application contexts and related ANR projects

I was involved in two different ANR (French National Research Agency) projects, both relative to securization and image coding topics. These projects, namely TSAR and CAIMAN deeply influenced my research when considering the content protection domain. The application contexts, inherent to these collaborations, have raised some concrete issues that we tended to alleviate. In this section, I briefly describe the purposes of TSAR and CAIMAN projects and my associated involvement.

TSAR project: Safe Transfer of high Resolution Art images - 2006/2009. The protection of digitized works of art still pose security problems when they broadcasted on-line. The goal of the project TSAR is then to transmit in a secure way high quality images. Within the TSAR project, five laboratories were involved, namely the C2RMF (Louvre, Paris), IETR (Rennes), IRCCyN (Nantes), LIRMM (Montpellier) and LIS (Grenoble).

Museums are supposed to undertake at least two essential missions [31]. First, they have to preserve their huge number of items and save them from damage. At the same time, museums play an active role in the spread of cultural knowledge and this educational objective leads them to widely communicate these materials. However, these two missions are somewhat contrary in nature because handling art items inevitably causes damage. To solve this major problem, museum research centers have introduced the "digital museum" concept [116]: digital versions of the original art items are collected in a database on a server accessible via the Internet.

For example, the National Gallery in London, the Tokyo University Digital Museum (through the Digital Museum 2000 project [167]) or the Chinese University Museum Grid [31] provide public access to their databases. However, users can only download low-resolution images. The application has been actually designed to prevent illegal copies of digitized data. The best way to achieve this consists basically of not transmitting high-resolution images. High-resolution information is therefore stored separately and reserved for a limited number of people.

In France, the C2RMF laboratory, connected to the Louvre museum, has digitized more than $300,000$ documents taken from French museums, in high resolution (up to $20,000 \times 30,000$ pixels). The resulting EROS database [149] is for the moment only accessible to researchers whose work is connected with the C2RMF. To widely open the database, the idea is to create a framework to integrate the different security solutions in order to secure the access to the images. This project then aims at transmitting hidden data in images of important sizes by means of steganography and watermarking techniques. This data hiding scheme will also be combined with cryptography and

compression methods in order to both increase the transfer security and to minimize the transmission time. The application concerns secure transmission of high-resolution images of painting and archaeological objects (several gigabytes of data).

As far as the French TSAR project is concerned, the Interleaved S+P image coding method takes part into the framework. The objective is to design an art image database accessible through a client-server process that includes and combines a hierarchical description of images and dedicated content securization processes. Jean Motsch, who I co-supervised, has defended his dissertation in 2009 in connection with the TSAR project. I was nominated as the scientific responsible of IETR works.

CAIMAN project: Codage Avancé d'IMAges et Nouveaux services - 2009/2012. In collaboration with Thalès Communications, and ETIS, XLIM-SIC, and IETR laboratories, the CAIMAN project aims at providing new image coders able to reach a widespread use such as JPEG does, together with advanced functionalities. In particular, a user-friendly solution should be designed, while keeping a low computational complexity and more advanced features in the direction of some tools of JPEG2K.

Not only the compression efficiency will drive the adoption of the future generation of image codecs but more the new services it will provide. These considerations have been taken into account by the JPEG Committee for JPEG advanced Image Coding to which CAIMAN will contribute. This CAIMAN project is thus naturally closely linked to JPEG-AIC work (see section 2.1).

The main objective of CAIMAN consists of studying a still image coder that jointly integrates in its design security aspects such as steganography, error-resilience, adaptation to the network and robustness to losses and Quality of Experience issues, without sacrifying to compression efficiency. The Interleaved S+P framework has been then proposed to the JPEG committee and related work led to contributions to the JPEG-AIC standard.

In particular, medical image concerns were addressed during this project. The Interleaved S+p coding scheme has been developed to face the secure transmission issues. Embedded functionalities such as adapted selective cryptography, human vision-based steganography coupled with Unequal Error Protection and error resilience tools have been designed so that to maintain good coding properties together with embedded Quality Of Service oriented system.

MPEG4-SVC video coding standard. The standard MPEG4-SVC was designed so that to add scalability. Technicolor took part to the standardization process of the framework, based in MPEG4-AVC coding scheme. However, the complexity of the proposed solution prevent from a widespread use of it. One of the identified bottleneck relies in the rate control mechanism. In collaboration with Technicolor, we provide an efficient one-pass rate control solution. Yohann Pitrey has then designed during his PhD work a low-complex solution, able to address both AVC and SVC frameworks.

4.2 Content protection features: Interleaved S+P application

Whatever the storage or channel transmission used, medical applications require secure transmission of patient data. Embedding them in an invisible way within the image itself remains a relevant solution. We also deal with security concerns by encrypting the inserted data. Whereas the embedding

scheme can be made public, the use of a decryption key will be mandatory to decipher the inserted data.

4.2.1 Steganography and the Interleaved S+P

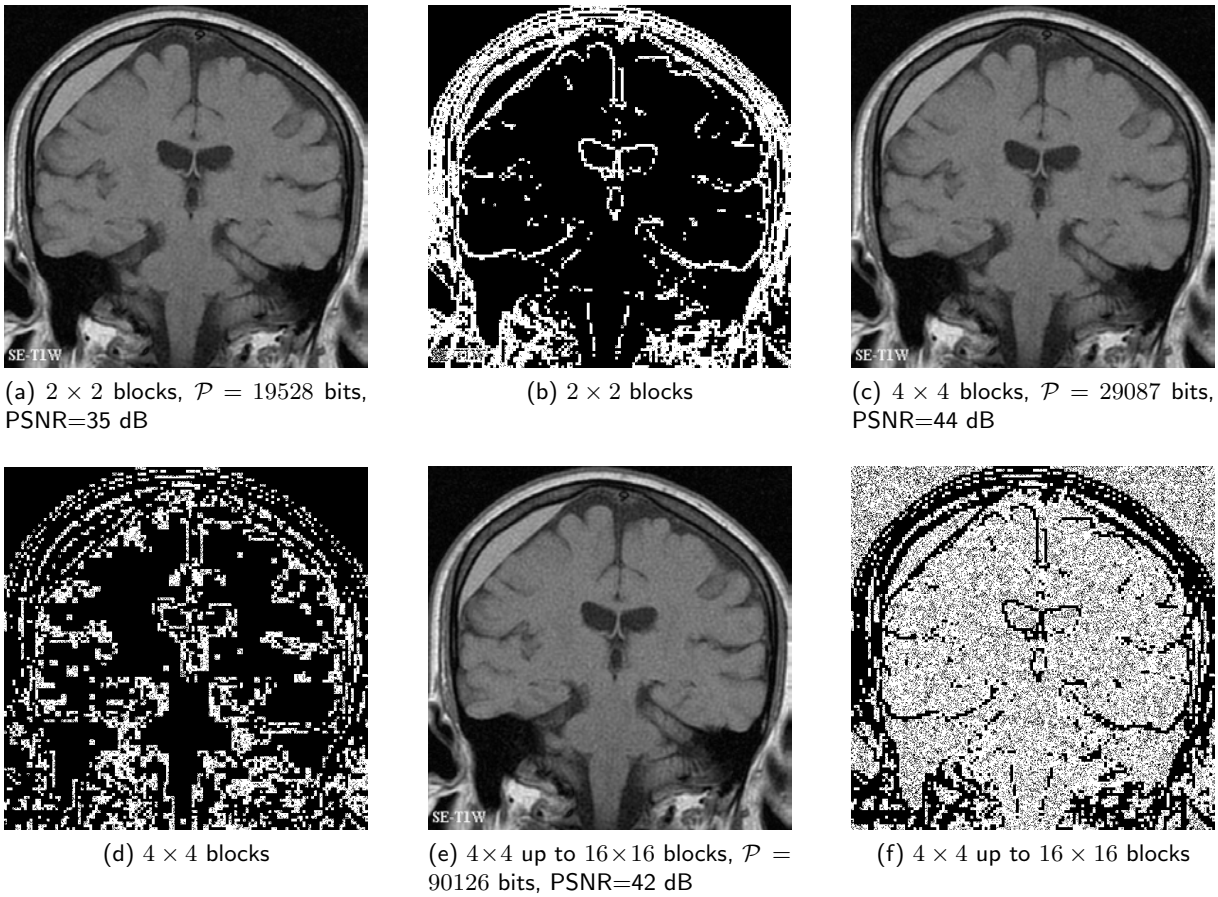
Data embedding is one of the new services expected within the framework of medical image compression. It consists of hiding data (payload) in a cover image. Applications of data embedding range from steganography to metadata insertion. They differ in the amount of data to be inserted and the degree of robustness to hacking.

From a signal processing point of view, it uses the image as a communication channel to transmit data. The capacity of the channel for a specific embedding scheme gives the size of the payload that can be inserted. A fine balance has to be achieved between this payload and the artifacts introduced in the image. This being so, different embedding schemes are compared on a payload versus PSNR basis. Of course, the overall visual quality can be assessed. The target application is the storage of data related to a given medical image or sensitive art images. That data can consist of patient ID, time stamps, or the medical report, transcribed or in audio form in case of medical materials. The idea is to avoid having to store several files about specific images by having all the necessary information directly stored within the image data.

We therefore propose a data embedding service that aims to insert a high payload in an image seen either as a cover or a carrier, such as a medical report in audio form. For this purpose, audio data, after coding and ciphering, is inserted in a corresponding image. The embedded image is then transmitted using usual channels. Of course, this scheme is compliant with any error protection framework that might be used. When retrieval of audio data is requested, the data embedding scheme is reversed, and both the original image and the audio data are losslessly recovered. To avoid significant perceptually distortions, the data hiding mapping is powered by the quadtree: distortions are less perceptible in homogeneous areas than upon edges as shown in figure 4.1.

In this context, we studied the Difference Expansion (DE) method, introduced by Tian [191] that embeds one bit per pixel pair based on S Transform. As the Interleaved S+P algorithm and DE both use S-Transform during their computation, we have combined both techniques to perform the data insertion without degrading coding performance. In order to adjust the DE algorithm to LAR Interleaved S+P, some minor modifications are introduced compared with the original DE method. In particular, we drive the insertion process by the quadtree partition, which means that the insertion is dependent on the image content. Another important improvement is that in the initial DE method, positions of possible "extensible" difference have to be encoded, adding a significant overhead. In our coding scheme, these positions can be directly deduced from the quadtree, and are then not transmitted [131].

We show preliminary results on an angiography 512-squared medical image (Figure 4.2). A payload of 63598 bits is inserted, with a PSNR of 40 dB. Considering a 1 million pixel image, the payload can be up to 300 kbits. Our embedding scheme performs well, allowing high payload and minimum distortion, as shown on zoomed parts of the images from the figure 4.2. From a compression point of view, the data hiding process does not affect the coding efficiency: the total coding cost is about equal to the initial lossless encoding cost of the source image plus the inserted payload.



(a) 2×2 blocks, $\mathcal{P} = 19528$ bits, PSNR=35 dB

(b) 2×2 blocks

(c) 4×4 blocks, $\mathcal{P} = 29087$ bits, PSNR=44 dB

(d) 4×4 blocks

(e) 4×4 up to 16×16 blocks, $\mathcal{P} = 90126$ bits, PSNR=42 dB

(f) 4×4 up to 16×16 blocks

Figure 4.1: Visual quality versus watermarked block sizes. For each image, position of modified pixels has been extracted (in white onto black background).

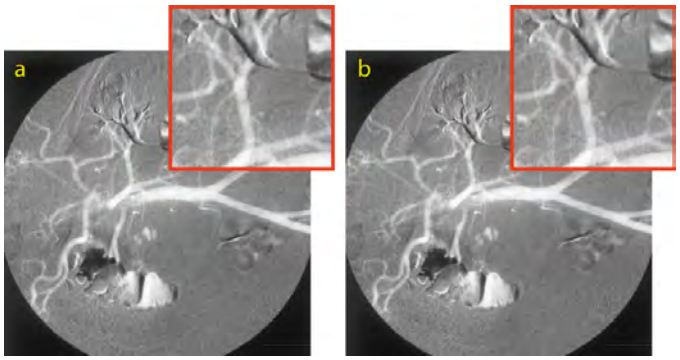


Figure 4.2: a) Source image - b) Image with inserted payload

4.2.2 Cryptography ans scalability

Besides watermarking, steganography, and techniques for assessing data integrity and authenticity, the provision of confidentiality and privacy for visual data is one of the most important topics in

the area of multimedia security in the medical field. Image encryption lies somewhere between data encryption and image coding. Specifically, as the amount of data to be considered is several orders of magnitude greater than the amount for ordinary data, more challenges are to be dealt with. The main challenge is the encryption speed, which can be a bottleneck for some applications in terms of computation time or in terms of required computing resources. A secondary challenge is to maintain the compliance of the encrypted bitstream with the chosen image standard used to compress it.

Partial encryption. Partial encryption addresses the first aforementioned challenge. Our partial encryption scheme is based mainly on the following idea: the quadtree used to partition the image is necessary to rebuild the image. This has been backed up by theoretical and experimental work. As a result, the quadtree partition can be considered to be the key itself, and there is no need to encrypt the remaining bitstream.

The key obtained is thus as long as usual encryption key and its security has been shown to be good. If further security is requested, the quadtree partition can be ciphered using a public encryption scheme, to avoid the transmission of an encryption key, as depicted in Figure 4.3 [132]. This system has the following properties: it is embedded in the original bit-stream at no cost, and allows for multilevel access authorization combined with a state-of-the-art still picture codec. Multilevel quadtree decomposition provides a natural way to select the quality of the decoded picture.

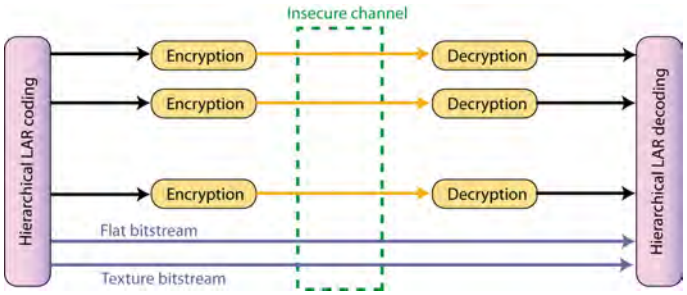
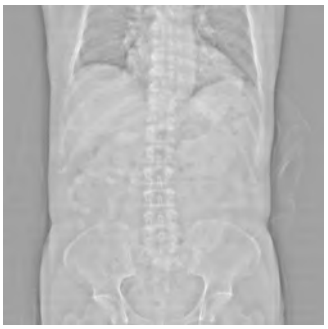


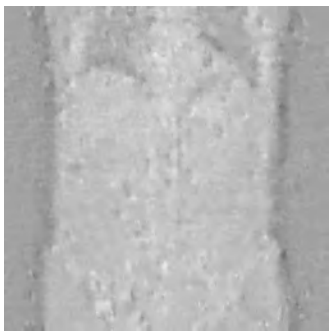
Figure 4.3: Interleaved S+P hierarchical selective encryption principle

Selective encryption. Selective encryption goes a bit further than partial encryption. The idea is to cipher only a small fraction of the bitstream, namely the main component, which gives the advantage of providing a decoder compliant bitstream. This property allows the user to see a picture even without the key. Of course, the picture must be as different to the original one as possible.

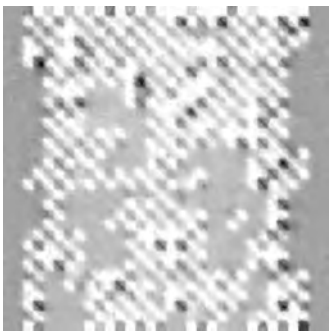
Our selective encryption scheme uses also the quadtree partition as a basis [57]. The data required in the compression framework to build the Flat picture are also used. The general idea is to encrypt several levels of the pyramid. The process begins at the bottom of the pyramid. Depending on the depth of the encryption, the quality of the image rebuilt without the encryption key varies. The encryption itself is performed by a well-known secure data encryption scheme. One main property of our selective encryption scheme is that the level of encryption (i.e. the level of the details remaining visible to the viewer) can be fully customized. Hierarchical image encryption is obtained by deciding which level will be decrypted by supplying only the keys corresponding to those levels. This refines the quality of the image given to different categories of viewers. The encryption of a given level of the partition prevents the recovery of any additional visually-significant data (figure 4.4). From a distortion point of view, it appears that encrypting higher levels (smaller blocks) increases the PSNR,



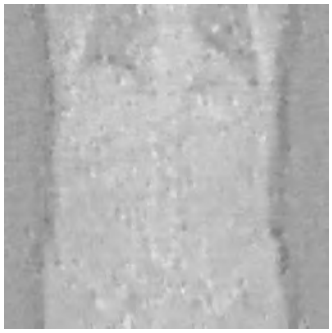
(a) Original image



(b) RC4 - Encryption of Quadtree partition



(c) RC4 - Encryption of Quadtree partition and Flat stream



(d) AES - Encryption of Quadtree partition

(e) AES - Encryption of Quadtree partition and Flat stream

Figure 4.4: Visual comparison between original image and image obtained from partially encrypted LAR encoded streams without encryption key.

and at the same time, the encrypting cost. From a security point of view, as the level increases, the search space for a brute force attack increases drastically.

4.2.3 Client-server application and hierarchical access policy

Images and videos databases are a powerful collaborative tool. However, the main concern when considering these applications lies in the secure accessing of images. The objective is therefore to design a medical image database accessible through a client-server process that includes and combines a hierarchical description of images and a hierarchical secured access.

Along with the TSAR project, a corresponding client-server application [11] has been then designed. Every client will be authorized to browse the low-resolution image database and the server application will verify the user access level for each image and eventual Region of Interest (ROI) request. ROIs can be encrypted or not, depending on the security level required.

If a client application sends a request that does not match the user access level, the server application will reduce the image resolution according to access policy. The exchange protocol is depicted in Figure 4.5.

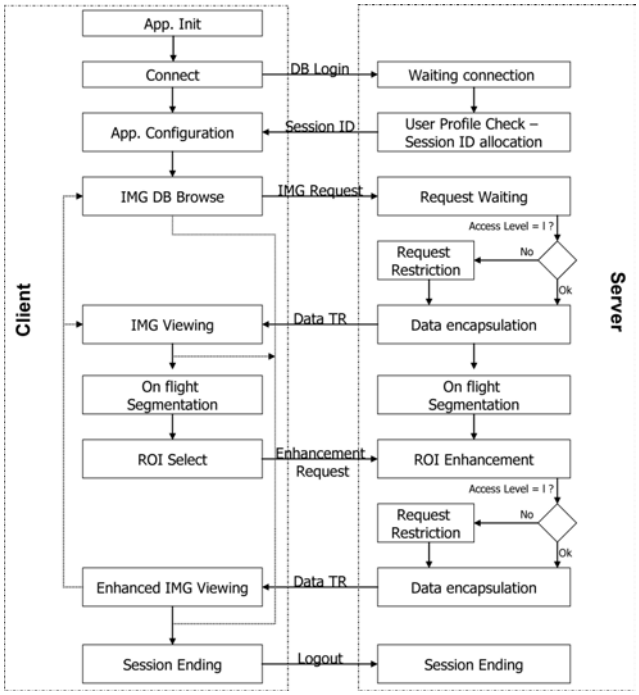


Figure 4.5: Exchange protocol for client-server application

4.3 Network oriented QoS solutions

If image content issues have been discussed in previous section, data integrity can also be handled by network oriented QoS solutions. Anyway, Quality of Services is a generic term gathering all kind of processes able to guarantee end-to-end requirements, such as final image quality, latency, frame rate, etc... In case of video coders, Rate/Distorsion allocation is naturally a part of QoS issues. In this context, PhD work of Yohann Pitrey was devoted to the design of a low complex rate control scheme targeted to SVC framework [147]. This work was realized in collaboration with Jérôme Viéron (Technicolor).

Besides, to ensure optimal visualization of transmitted images, there are two possible ways of protecting the bitstreams. First, protecting the encoded bit-stream against error transmission is required when using networks with no guaranteed quality of service (QoS). In particular, the availability of the information can be ensured by the Internet protocol (IP). We focused our studies on two topics, namely the loss of entire IP packets and the transmission over wireless channel. When considering the Interleaved S+P coder, we develop error resilience strategies adapted to our compression scheme. In particular UEP solutions used together with proper resynchronization processes and robust encoding naturally leads to optimal conditions for the transmission of sensitive data. Even if we also developed operative resynchronization processes [23] relying on Q15 like coder (section 3.4), only information about UEP framework will be given in this part.

4.3.1 One-pass rate control scheme using ρ-domain for MPEG-4 SVC

Video coding and processing have become central areas of interest for video communications. Video coding standards such as MPEG-4 AVC/H.264 [206] manage to efficiently reduce the amount of data to transmit. However, current video applications have to cope with heterogeneous communication networks and video-reading devices. Conventional video coding suffers from the lack of adaptability to these multiple targets. Typically, a different version of the video must be encoded for each target. Moreover, the redundancies between the different versions of the video are not exploited, which results in a waste of time, storage space and bandwidth.

As a response, scalable video coding has been developed to cope with this need for adaptability. Recently a standard was supplied, with the finalization of the new MPEG-4 Scalable Video Coding extension (SVC) [156, 179]. This standard supports three types of scalability. Spatial scalability affects frame resolution, to address variable screen sizes. Quality scalability acts on the signal-to-noise-ratio (SNR) to provide various levels of quality in the decoded video stream. Temporal scalability increases the number of frames per second to improve motion smoothness. An MPEG-4 AVC/H.264 compatible base layer is then first encoded, then motion and texture information can be upsampled and refined to code the enhancement layers more efficiently. This new tool called inter-layer prediction enables MPEG-4 SVC to provide scalability features without losing too much coding efficiency when compared to MPEG-4 AVC/H.264 [171].

MPEG-4 SVC ensures that a video stream can adapt to different decoding targets. However, the standard does not provide any tools to cope with the communication channel capacity. Activity variations in the contents of a video sequence can cause great bitrate fluctuations at the output from the encoder. If not controlled, such fluctuations can cause undesirable display interruptions in the decoder. To transmit an encoded video stream on a communication channel, the bitrate of the stream must cope with the channel bandwidth. Rate control is a critical part of the encoding process, as it is intended to regulate the bitrate and attenuate these fluctuations. Generally, a budget is first determined according to the available bandwidth, and dispatched among the video frames. Then, a bitrate model anticipates the behavior of the output bitrate from the value of the quantization parameter, in order to reach the desired budget. Although rate control has been widely studied for conventional video coding such as MPEG-4 AVC/H.264 [115, 107, 72], only a few proposals exist for scalable video coding. In [1], the rate control scheme from [107] is implemented in MPEG-4 SVC but only affects the base layer. In [214], only spatial and quality scalabilities are handled and each macroblock is encoded twice, which dramatically increases the computational complexity of the encoder. The scheme presented in [110] is able to control each type of scalability in SVC using only one encoding pass. The statistics from the base layer and from the previous frame in the same layer are used to predict bitrate behavior and distorsion. Although these contributions are of great interest for rate control on MPEG-4 SVC, they remain quite complex and require a lot of calculations. Moreover the tested configurations do not reflect practical SVC applications, and do not cope with current video stream bitrate and size requirements, as specified in [84].

During his PhD work, Yohann Pitrey has to design a new one-pass low-complexity rate control framework, dedicated to this video codec. We control the bitrate at frame level using a low-complexity linear bitrate model based on the ρ-domain framework [72]. This model is used to process the quantization parameter of each frame. In particular, we use the statistics from the previous frame to initialize the bitrate model. This way, each frame is only encoded once and the computational complexity of the whole rate control process is extremely low. Additionally, we dispatch the target

bitrate inside each group of pictures to get smooth quality variations in the encoded stream. Frame weights based on their coding efficiency are used to reduce the PSNR variations in each group of pictures.

4.3.1.1 ρ-domain based rate model

Conventional rate control approaches try to formulate the bitrate as a function of the quantization parameter (QP) [158, 115]. Indeed, the QP determines the amount of data lost during the encoding process and has a direct influence on the output bitrate. Based on this relationship, it is possible to choose the optimal value of QP, by predicting its impact on the output bitrate. However, the relationship between the QP and the bitrate is difficult to approximate correctly. To alleviate this problem, a common approach is to encode each image with several values of QP and choose the value that produces the bitrate closest to the constraint. This kind of exhaustive approach is not suitable in practice as it requires a lot of computations. Other approaches try to estimate the distribution of the data before quantization using Laplacian or Gaussian functions [158, 108]. This estimation is then used to predict the behavior of the bitrate from the value of QP. Unfortunately, the approximation step remains quite complex and suffers from inaccuracies.

Another approach called ρ-domain uses the amount of null coefficients in a frame after quantization as an intermediate parameter between the QP and the bitrate [72]. It has been observed that this parameter, denoted as ρ, has a direct influence on the bitrate needed to code a frame [73]. The relationship between ρ and the bitrate is highly linear, which makes it easy to evaluate [72]. A relationship can be found between ρ and the QP, to relate the bitrate to the QP. In [175], the so-called ρ-domain rate model was used to successfully control the bitrate on MPEG-4 AVC/H.264.

4.3.1.2 The ρ-domain model for MPEG-4 AVC

After the prediction step, the residual information is transformed using an Integer Discrete Cosine Transform (IDCT). The transformed coefficients are then quantized and sent to entropy coding. Considering the DCT coefficients of a frame, it is easy to determine how many of them will be coded as zeros after quantization. This coefficient will be coded as a zero if its value is below a specific dead zone threshold. In the quantization scheme used in MPEG-4 AVC, the threshold depends on both the position of the coefficient in the transformed macroblock and the value of the QP [175].

In [175], it is stated that ρ can be expressed as a function of bitrate R as follows:

$$\rho(R) = \frac{R_0 - R \times (1 - \rho_0)}{R_0}, \tag{4.1}$$

where R_0 and ρ_0 are two initial values to be determined. It is obvious that this relationship is linear. Note that the couple of values $(R; \rho) = (0; 1)$ is a solution of equation (4.1). Indeed, when the bitrate is equal to zero, all the coefficients are coded as zeros, so ρ is equal to 1.We can then find the value of QP q_t that generates the closest number of bits to a target bitrate R_t, so that

$$q_t = \underset{q \in [0;51]}{\arg\min} \left| \rho(q) - \rho(R_t) \right|. \tag{4.2}$$

The ρ-domain modeling framework has several advantages. First, this model is very accurate and no approximation are required, as the quantized coefficients are directly available during the

encoding process. Secondly, the ρ-domain has very low computational complexity, as the model in equation (4.1) is linear.

In [148], we validated the ρ-domain model for MPEG-4 SVC purposes. Then, model initialization issue has to be dealt.

4.3.1.3 Initialization of the ρ-domain model

Two approaches have been proposed. In our first step work [148], such as in [72, 175], each frame or macroblock is pre-encoded to get the quantized coefficients. These coefficients are then used to process the value of ρ and choose the right value of QP, which is used in a second encoding pass. This so-called *two-passes* solution achieves good results, as the rate model is initialized with data from the frame itself. However, the encoding process is executed twice and the computation time is significantly increased. We then propose to use the information from the previous frame as a basis for the calculation of ρ. This kind of approach has already been studied in other bitrate representation contexts [107, 217]. Spatial and temporal correlations between consecutive frames are used to predict the coding parameters. The main advantage of this so-called *one-pass* alternative is that no pre-encoding step is needed, thus leading to lower complexity.

The next section presents a one-pass rate control scheme for MPEG-4 SVC using the ρ-domain model.

4.3.1.4 Global rate control strategy.

Our scheme was designed as a compromise between low computational complexity and accurate rate control. For this reason, we execute the rate control step only once per frame and use the same QP for the whole frame. A drawback of this is that we generally cannot reach the exact budget for a frame. Changing the QPs for all macroblocks at a time induces a threshold effect between the reachable bitrates at the frame level. Some existing rate control approaches run at the macroblock level. They manage to control the bitrate more accurately, but at the cost of higher computational complexity.

The quality of user visual experience is closely related to the quality variations inside the decoded video stream. Thus, we aim to reduce the PSNR fluctuations inside each GOP. To do this, we use a frame type dependant budget allocation to dispatch bits among frames according to their coding complexity.

Budget allocation. Budget allocation is an important part of a rate control scheme. Most of the choices are made at this stage, and the QP processing module is designed exclusively to respect these choices. In most practical SVC applications, each layer addresses a particular target, with specific bitrate requirements. So we specify a bits-per-second (bps) constraint for each layer, which is handled separately. This target bitrate is then converted into bit budgets at GOP and frame levels.

GOP-level budget allocation The available bitrate is first dispatched among GOPs. Inside a given layer l, each GOP is granted the same budget. G_l defined as

$$G_l = S_l \times \frac{C_l}{F_l} + E, \tag{4.3}$$

where C_l is the required target bitrate per second, S_l is the size of a GOP in layer l and F_l is the number of frames per second in layer l. We add a small feedback term E to compensate for allocation errors from previous GOPs. In our experiments, E is limited to 10% of the entire GOP budget.

Once GOP level budget allocation is completed, we have to dispatch the budget among frames. Great care must be taken in dispatching the budget among the different types of frame, because it has a direct impact on output quality. To this end, we define the following frame weights.

Relative frame-weights MPEG-4 SVC supports three types of frames (*i.e.:* I, P and hierarchical B-frames). Each type uses specific coding tools and has distinct coding performances. I frames use only intra-frame macroblock prediction and are the most reliable. However, their coding efficiency is not very high. P frames allow intra and inter-frame prediction and benefit from better coding efficiency than I frames. B frames use bidirectional inter-frame prediction and are the most effective. As a result, getting the same quality requires more bits for P frames than for B frames. MPEG-4 SVC also provides a hierarchical GOP structure using hierarchical B frames to ensure temporal scalability [170]. Successive B frames are encoded in a pyramidal fashion so that when a level is added, the number of frames per second is multiplied by two. This GOP structure causes the coding performances of hierarchical B frames to vary depending on their temporal level. MPEG-4 SVC allows eight temporal levels for B frames. We consider each temporal level as a different frame type, namely $B_1, B_2, \dots B_8$.

To ensure constant quality within a GOP, we dispatch the allocated budget according to the coding performances of each type of frame. For each type of frame T in each scalable layer l, we introduce a frame weight $K_{T,l}$, defined as

$$K_{T_f,l} = 2^{q/6} \times b_f, \tag{4.4}$$

where T_f is the frame type of frame f, b_f the number of bits needed to code the frame f, and q the QP value. As it depends both on the QP and the bitrate, this weight reflects the coding efficiency achieved for a frame. Basically, with equal QP, $K_{I,l} > K_{P,l} > K_{B_1,l} > \cdots > K_{B_8,l}$. To obtain constant quality inside a GOP, we dispatch the available budget among frames according to each frame type's weight.

Frame-level budget allocation. We use then the frame weights to dispatch the GOP budget among frames. A frame needs to be allocated a budget that corresponds to its relative weight in the GOP. The target budget R_t for a frame at position f in a GOP in layer l is processed as follows:

$$R_t = \frac{\tilde{K}_{T_f,l}}{\sum_{i=0}^{f-1} K_{T_i,l} + \sum_{i=f}^{S_l} \tilde{K}_{T_i,l}} \times G_l + \epsilon, \tag{4.5}$$

where T_i is the type of frame at position i and ϵ is a small feedback term to compensate for the allocation errors from previous frames. As rate control is processed before encoding frame f, we do not know its weight. As an estimation, we use $\tilde{K}_{T_f,l}$ which is the weight of the last encoded frame that has the same type in the same layer. Similarly, we use the last encoded frame weights $\tilde{K}_{T_i,l}$ for frames that have not yet been encoded.

QP processing Once the budget allocation step is complete, we choose the optimal value of QP for each frame. The optimal target value of QP, denoted as q_t, is the one that minimizes the difference between the number of bits needed to encode the frame and the target number of bits, denoted as R_t. Using the one-pass approach, the target value of ρ is processed as

$$\rho_t = \frac{R_p - R_t \times (1 - \rho_p)}{R_p}, \tag{4.6}$$

where q_p is the value of QP used to encoded the previous frame of the same type in the same layer, R_p is the generated number of bits and ρ_p is the corresponding value of ρ. Then, the target QP is determined as

$$q_t = \underset{q_p - \Delta_q \leq q \leq q_p + \Delta_q}{\arg\min} |\rho(q) - \rho_t| . \tag{4.7}$$

4.3.1.5 Experimental results

We define three scenarios to test spatial, quality and temporal scalabilities. The tested configurations are summed up in table 4.1. All our tests were based on the JSVM Reference Software version 8.6 [179] in which we implemented our ρ-domain-based rate control algorithm. We use reference video sequences to attest the performances on various types of contents.

Figure 4.6 shows the achieved bitrates per second for two video sequences. The achieved bitrates are very close to their targets, which shows the ability of our scheme to respect the specified constant bitrate per second constraints. We also display the behaviour of our scheme at frame level in figure 4.7. It shows the accuracy of the bitrate control at frame level, as the achieved bitrate is also very close to the constraint. Additionally, the behaviour of the budget dispatching method is clearly visible. P frames are granted more bits than the B frames, and the B frames get less and less bits when their temporal level increases. The impact on the overall quality will be discussed with more details in the following. The next evaluates the proposed scheme in terms of encoding time.

Our budget dispatching policy grants a number of bits to each frame according to their relative coding complexity. The differences between frame types and levels are therefore compensated and the quality along the GOP should be constant. Figure 4.7 displays the achieved bitrate at frame level. Using our budget dispatching policy, P frames are granted more bits than B frames, while B frames are granted more bits in the lower levels than in the higher levels. We compare the results

Scenario	Layer	frame size (pixels)	frames per second	target bitrate (kbps)	frames per GOP
SPATIAL	0	176*144	30	100	16
	1	352*288	30	400	16
	2	704*576	30	1200	16
QUALITY	0	352*288	30	400	16
	1	352*288	30	1200	16
	2	352*288	30	3600	16
TEMPORAL	0	352*288	15	200	4
	1	352*288	30	400	8
	2	352*288	60	800	16

Table 4.1: Test scenarios for each type of scalability.

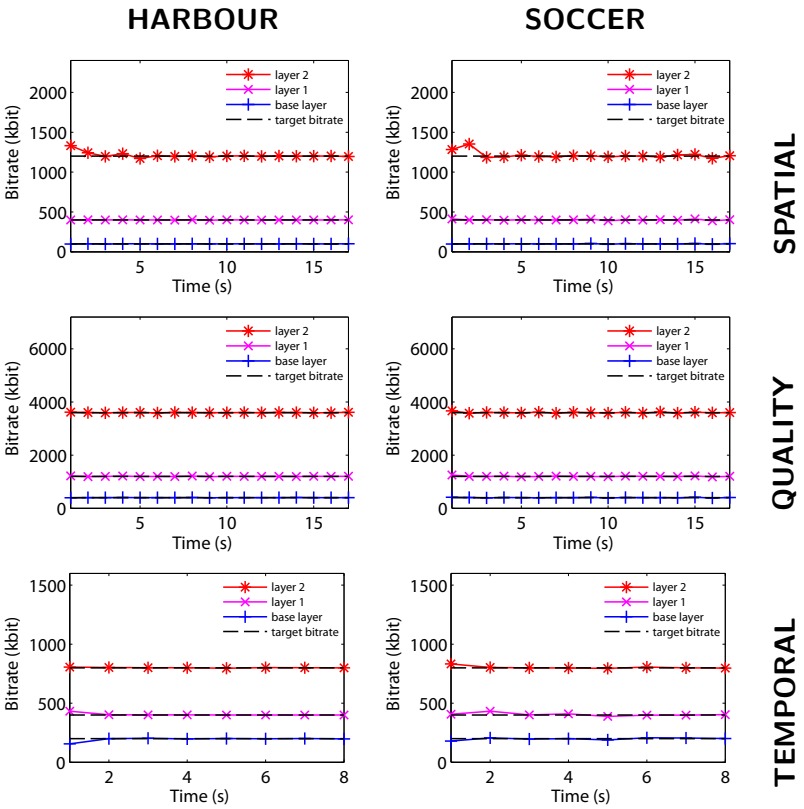


Figure 4.6: Achieved bitrate per second using our one-pass rate control scheme.

obtained with this policy to a simple policy that grants the same number of bits to each frame, or so-called "constant-budget" policy. Visually, the constant-budget policy produces quality fading effects between the P frames, which are quite unpleasant for the viewer. Using our policy, the quality is smoother and the overall quality impression is much better. It is interesting to note that we also obtain a slight increase in PSNR. Actually, the quality of P frames is higher, so they make a better prediction for B frames. As a consequence, the quality of the whole encoding process is slightly higher using our dispatching policy.

As a result, the presented one-pass rate control scheme performs accurate bitrate control, while successfully reducing the quality variations along the video frames.

4.3.2 Error resilience and UEP strategies

Commonly, robust wireless transmission can be achieved through the use of error resilience processes at both source and channel coding. At the source coding side, the entropy coder is often the less robust part of the coder. When using arithmetic entropy coder such as MQ coder used in JPEG2K format, a single bitshift in the bitstream is enough to create important visual artifacts at the decoding side.

First, to prevent the decoder from desynchronizing and therefore error from propagating, synchronisation markers needs to be added. Moreover, specific error detection markers can be used to

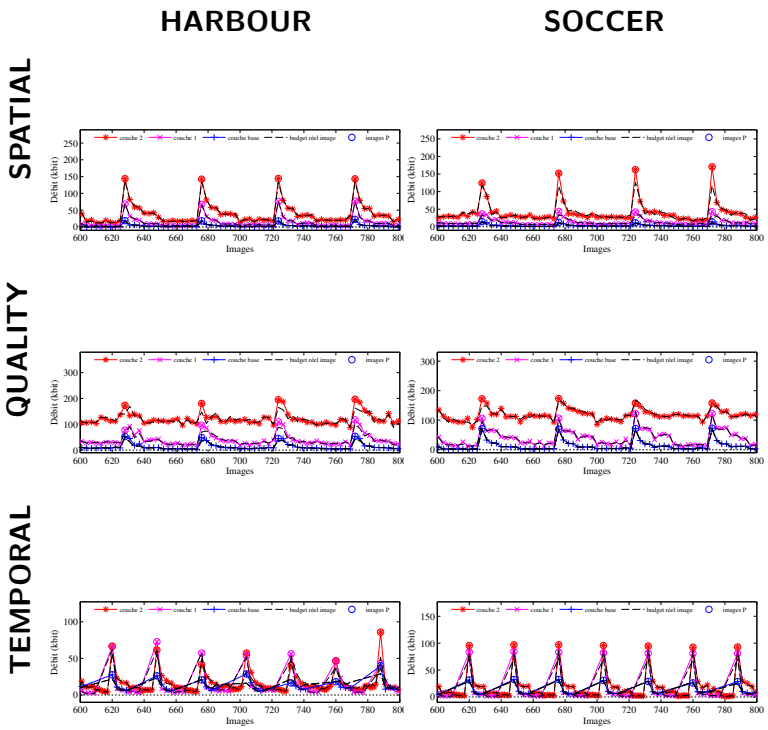


Figure 4.7: Achieved bitrates at frame level using our one-pass rate control scheme.

detect errors during decoding and discard bitstreams affected by this error. Such synchronization and error detection markers have already been implemented as SEGMARK, ERTERM and RESTART markers in the JPEG2K codec [189] as well as in the Interleaved S+P codec [142].

At the channel coding side, error robustness is achieved by using error correcting codes, such as Reed Solomon [155] and convolutive codes [52]. Such error correcting codes add redundant data in the bitstream in order to detect and possibly correct transmission errors. Depending on the channel characteristics, error correcting codes have to be tuned to achieve good performance in error correction while keeping a small amount of redundant data. These error correcting codes are usually computationally expensive and fast codes like LDPC and turbo codes can often be used instead.

Moreover, bitstream overhead has to be taken in consideration while performing image compression and has to remain as low as possible to maintain an acceptable bit rate. In this section, we only consider IP packet-based error correcting codes.

4.3.2.1 IP packets securization processes

Very few works cover the loss of entire IP packets in medical data transmissions [128] in particular. In a more general framework such as image transmission, most studies relate to the implementation of error control coding e.g. Reed-Solomon codes to compensate for packet loss by avoiding retransmissions [128, 47].

By adjusting the correction capacities and, thus, the rates of redundancy, it is possible to adapt to both a scalable source and an unreliable transmission channel. This is the purpose of Unequal Error Protection (UEP) codes which have been proposed within standardization processes [50]. The specific

problem of medical or art image integrity is the volume of the data being transmitted (cf lossless coding, 3D-4D acquisition etc.). Within this framework, UEP must meet algorithmic complexity requirements to satisfy real time constraints.

Roughly speaking, the most important part of the image is more protected by redundant information than non significant data. Figure 4.8 illustrates the associated framework. From image coding process, both bitstream data and codec properties are available for an advanced analysis stage. Then, a hierarchy can be extracted from the bitstream, so that the UEP strategy stage can add adequate redundancy. As a consequence, fine granularity can be obtained for good adaptation both to the hierarchy of the image and to the channel properties as joint source channel coding.

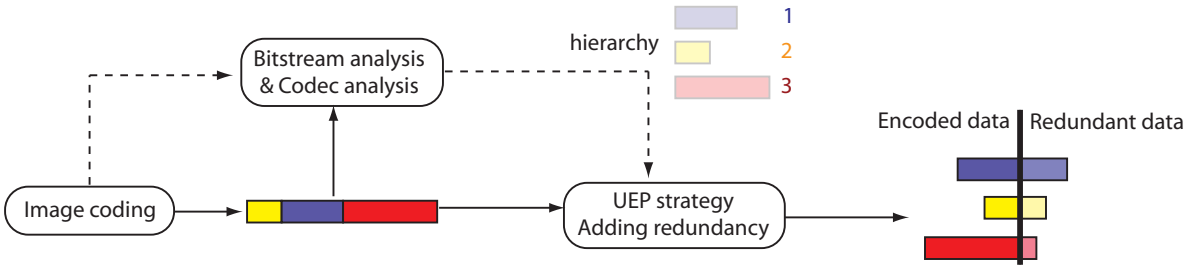


Figure 4.8: UEP principles: hierarchy and redundancy

A great deal of research work has been done in this area over the past decade. In particular, the working draft of JPEG2K WireLess (JPWL) [50] proposes to concentrate unequal protection on the main header and the tile header as any error on these part of the stream is fatal for decoding. In this solution, conventional Reed-Solomon error correction codes are applied to a symbol level to provide protection [47]. A very strong protection obviously improves the chances of success in decoding when binary losses occur but it also guarantees the integrity of the headers whether the properties of the channel are good or very bad. Furthermore, performance evaluation and protection on a symbol level are far removed from the real channels like wireless channels as can be seen for example through the variations in the protocol IEEE802.xx (WLAN or WiMax). More precisely, the approach never considers the effectiveness of the mechanisms operated on the level of Media Access Control (MAC) layer and physical (PHY) layer such as the Hybrid ARQ (Automatic Query Request - H-ARQ) combining efficient channel coding (turbo-code) and retransmission. Likewise, unequal error protection [2] or the new representations based on a multiple description of information [210] are not considered. Still, when designing joint source-channel coding UEP schemes, we jointly consider the PHY and MAC layers as effective to deliver true symbols so as to propose unequal protection framework at the transmission unit level *i.e* the packet level.

4.3.2.2 UEP strategy for scalable codec

When considering wireless channel, limited bandwidth and SNR are the main features. Therefore, both compression and secure transmission of sensitive data are simultaneously required. The Interleaved S+P and an Unequal Error Protection strategy are applied respectively to compress and protect the original image. The UEP strategy takes account of the sensitivity of the substreams requiring protection and then optimizes the redundancy rate. In our application, we used the Reed Solomon Error Correcting Code RS-ECC, mixed with symbol block interleaving for simulated transmission over the COST27 TU channel [71] (Figure 4.9). When compared to the JPWL system, we

show that the proposed layout is better than the JPWL system, especially transmission conditions are bad (SNR < 21 dB).

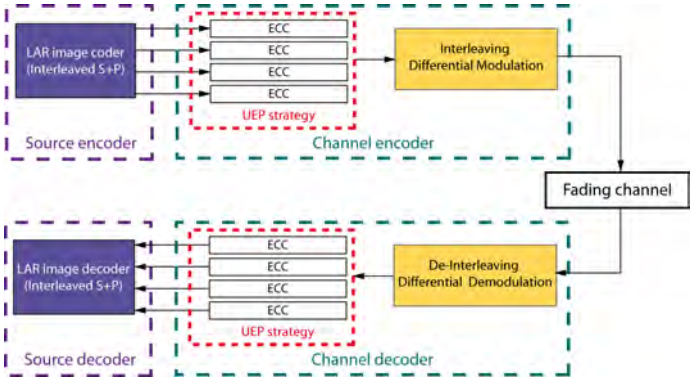


Figure 4.9: Overall layout of the multi-layer transmission/compression system

In other words, compensating IP packet loss also requires a UEP process, which uses an exact and discrete Radon transform, called the Mojette transform [22]. The frame-like definition of this transform allows redundancies that can be further used for QoS purposes (figure 4.10).

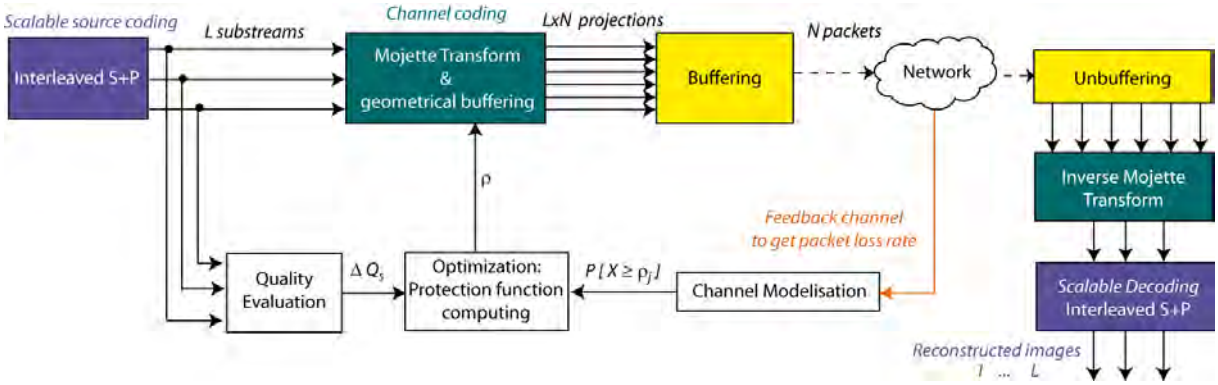


Figure 4.10: General joint LAR-Mojette coding scheme

Other simulation tests have been designed for MIMO systems using a similar framework, and have shown the ability of our codec to be easily adapted to bad transmission conditions, while keeping reasonable additional redundancy. At this point, comparisons with other methods remain difficult. Both SISO and MIMO transmissions simulation tools were provided by the French X-LIM Laboratory [27]. This work has been supported successively by the TSAR and the CAIMAN projects. Current developments are focused on the study of the LTE transmission system [126], and its combination with Interleaved S+P coded bitstreams.

These preliminary tests have been carried out without implementing basic error resilience features, such as resynchronization process, that should greatly improve our results. Some related solutions are presented below. Typically, resynchronization solutions based on QM-like entropy coding scheme (see section 3.4) drastically increases the overall performances of the system.

4.4 Conclusion

This chapter was dedicated to joint QoS oriented image coding and securization framework. Quadtree-based cryptography and data hiding were shown to be efficient solutions for content securization. In terms of error resilience, source-based together with channel-based coding have been developed.

MPEG4-SVC dedicated rate control has also shown is efficiency. Indeed, our rate control approach competes with the existing approaches, as the results can attest. The main advantages of our scheme are the accuracy of the ρ-domain bitrate model and its low computational complexity, as only one pass is performed. The presented bitrate dispatching policy also allows our method to obtain a smooth quality throughout each GOP, which enhances the perceived quality and insure end-to-end QoS based functionality.

As for it, the Interleaved S+P coding scheme has been developed to face the secure transmission issues. Embedded functionalities such as adapted selective cryptography, human vision-based steganography coupled with Unequal Error Protection and error resilience tools have been designed. The idea is to maintain good coding properties together with embedded Quality Of Service oriented system. This framework has yet been evaluated by the JPEG committee and has shown its global efficiency, even if the global framework has not been retained.

However, the exchange of medical data remains a key research topic. As for the moment, when considering medical image protection issues, Picture Archiving and Communication System (PACS) oriented frameworks have limitations in terms of securization process durability. If classical medical frameworks use image coding schemes such as JPEG, JPEG2K, JPEGXR, securization processes only act as additional features. A complete joint system should be built in such a manner that both coding and secure properties would benefit from each other. This remains an open research area.

Chapter 5

Generic analysis tools

For QoS/QoE purposes, generic coding tools as well as data integrity protection systems have been designed and described in previous chapters. To improve QoE, quality enhancement of images as well as content handling services can be additionally proposed. This chapter then deals with generic solutions for image analysis purposes. It can be then seen as a toolbox easily adaptable to more complete image processing chain.

First, image interpolation issue is addressed. As an example, resolution scalability of the Interleaved S+P codec is an important property as it embeds several versions of the image at different resolution levels. The resolution of these levels is doubling in size up to the original resolution and each resolution level is predicted from the previous one. In other words, the Interleaved S+P codec predicts, or interpolate, a bigger image from a smaller one. Our idea is to focus on this dyadic interpolation process and to propose a better and generic interpolation method. The underlying interpolation method should get the highest quality possible for a minimal complexity.

On the other hand, all images contain a certain level of semantic. Recognizing objects and naming them is such a complex task that most of approaches have been introduced at a pseudo semantic level. One approach is to extract coherent areas in the images that share uniform characteristics. These similar areas are then grouped together thanks to a segmentation process. Image segmentation creates regions where the semantic information is uniform regarding to some criteria. Even if these regions do not match real world semantic, the information can nevertheless be exploited. The semantic information can be then extracted in terms of computed abstract features such as texture, colors, shapes etc. The image can also be translated in a more abstract feature space to reduce the image characteristics to a small set of parameters. In this feature space, similar images have closed characteristics and are easier to compare or interpret. Depending on the application, semantic extraction process has various constraints and purposes. The major constraint is the algorithm complexity versus performance, the more the algorithm mimics the human brain the more time expensive it is.

Depending on the application, the segmentation algorithm can be tuned to reach a proper semantic level enabling a given task even if this pseudo semantic is nothing close to a real world semantic. In this context, most region segmentation methods focus on getting the best segmentation representation using various approaches [218]. The goal of those methods is to automatically extract the semantic information from an image to address various applications such as object recognition [69] or classification [114].

Region segmentation can be either performed directly in image domain [5] or in a feature space using clustering techniques [54]. The main drawback of most methods is the high complexity due to the computational complexity and the amount of data to process. In order to reduce the complexity, some suboptimal region segmentation methods first reduce the amount of input data leading to faster methods, for instance by using a quadtree partitioning [91]. In particular, quadtree structures are typically used in video [121] and image coding [209]. As an example the draft specification of HEVC video standard specifies variable block size coding units (CUs) coded through a quadtree structure. In this chapter, we then propose two segmentation solutions that take advantages of the quadtree-based image representation.

This chapter is organized as follows: first the Dyadic Fast Interpolation method is presented in section 5.1. Then quadtree based segmentation techniques are given. In section 5.2, we designed a solution that only takes the quadtree structure as input. It then can be seen as a kind of proof of concept, leading to underline the ability of the quadtree to represent image content. Then, a multiresolution segmentation is shown in section 5.3, relying on multiresolution Region Adjacency Graph structure.

5.1 Dyadic Fast Interpolation

Interpolation is required in applications when the original size of the image is not adapted to the desired usage. In this case, images must be resized to fit to the desired resolution and the most difficult task is image enlargement. Indeed, in case of enlargement, the subsequent interpolation process has to estimate missing pixel values. In the literature, interpolation is sometimes referred as single image super-resolution. A wide spectrum of approaches have been proposed for interpolation, together with a large range of complexity.

State-of-the-art interpolation methods can be classified in two distinct groups (figure 5.1). The first group (in green) contains high complexity methods with usually high visual quality. The second group (in blue) contains low complexity methods with often lower quality.

On the high complexity, high quality side, methods extract some statistical information from low resolution data to reconstruct high resolution images. Some approaches build high resolution images by stitching several high resolution patches [62] that locally correspond to low resolution image. This type of methods gives good results but is highly dependent on the patches database and tends to create unnatural artifacts. Some methods like [164] or [194] use Partial Differential Equations (PDEs) with excellent visual results, at the expense of high complexity. NEDI [106] and improved methods such as iNEDI [7] compute missing pixels by calculating the contribution of the neighbor pixels to the missing ones and assuming that covariance stays constant across scales. If this method is among the most popular in the field, its complexity is rather high (1300 multiplications per pixel). With quality comparable to NEDI [106], ICBI [61] uses second order image derivatives and iterative process leading to a moderate complexity. Another type of methods perform a local analysis to select the best appropriate oriented linear filter in a given filter bank. These approaches can be in particular based on wavelets [30] and lead to moderate complexity property. However, the resulting images are not as good as high-ends methods.

On the low complexity, lower quality side, methods are mostly based on linear kernels that are less computationally expensive but provide medium to low quality image. The best images are generally obtained by sinus cardinal (sinc) based kernels like Lanczos [46] filter. This type of filter

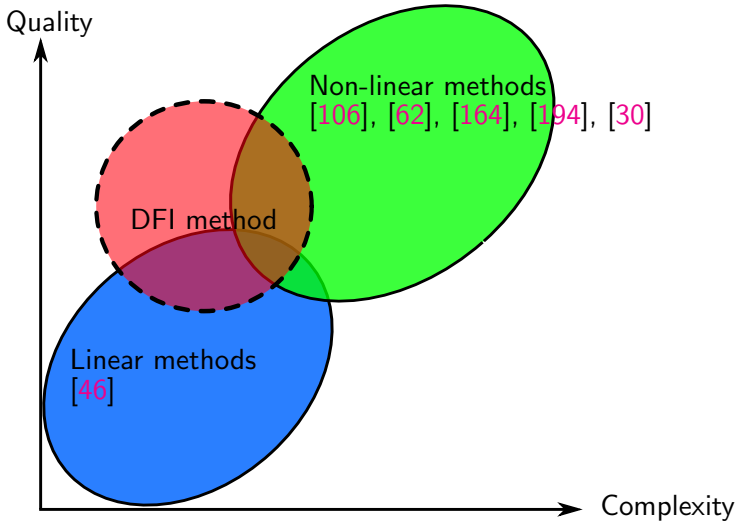


Figure 5.1: Complexity - quality target of presented DFI method

is theoretically ideal in the continuous domain according to the signal theory. Those kernel-based methods are the ones used in major image softwares because of their low complexity and flexibility. In particular, these methods can reach any output image size while some more complex method cannot such as NEDI [106]. The drawback of these simple methods remains the generated artifacts (jagger, blur) that can strongly penalize the visual results.

This section presents the Dyadic Fast Interpolation (DFI) method that can be applied in different contexts. The complexity and quality targets of DFI method is depicted in red on figure 5.1 in comparison with other methods. Our idea is to design a low complexity interpolation solution that provides visual quality as close to the state-of-the-art solutions as possible. This work has been elaborated by Clement Strauss in [182].

5.1.1 Geometric duality principle

What happens in images at pixel level is not completely chaotic and pixels can be predicted from their surrounding areas through "geometric regularity" of edges [118]. The proposed DFI method exploits more precisely the geometric duality properties in images. Geometric duality refers to the similarity found across scales: from one scale through another, pixel surrounding areas behave similarly. More precisely, edge behavior remains stable across scales and can be estimated from one to the other.

NEDI [106] and NEDI-based methods use this duality principle with covariance matrix: basically a local covariance matrix is computed at low resolution scale then is projected on the high resolution scale to compute the missing pixels. The main drawback of this technique is the high complexity involved by the computation of the inverse matrix.

The idea of DFI algorithm is to exploit the geometric duality in a simpler way. As shown in figure 5.2, the differences from the center green pixel to its four red neighbors are equivalent to the same pattern, slightly shifted, at a larger scale. The DFI method exploits the differences from one pixel to its neighbors. The duality property is used to retrieve unknown differences at a given scale from the

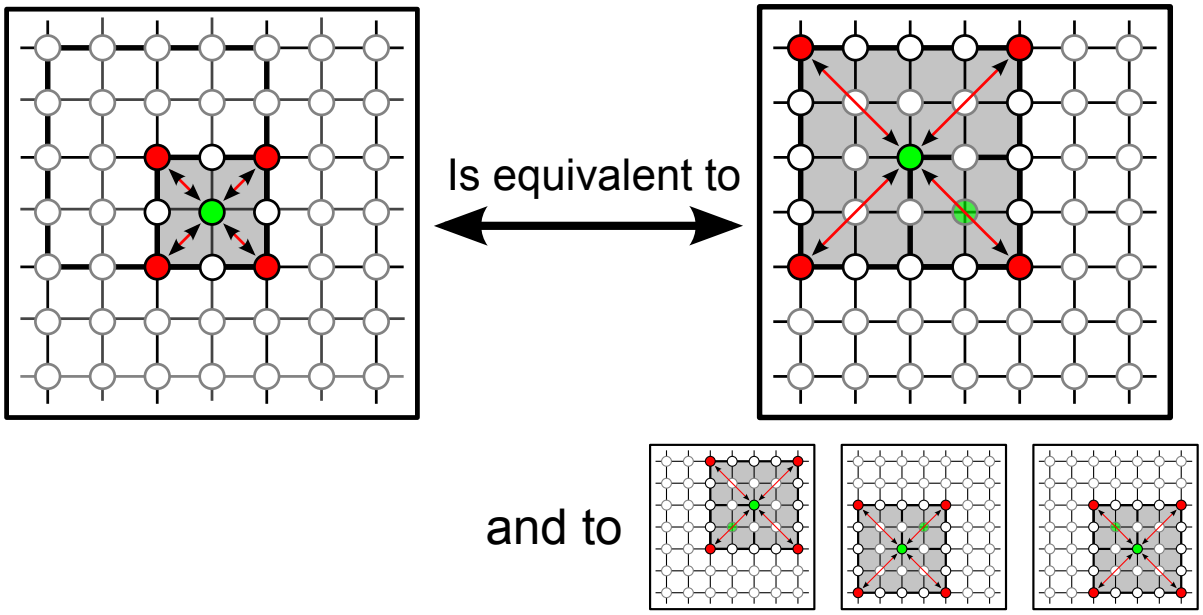


Figure 5.2: Geometric duality across scale

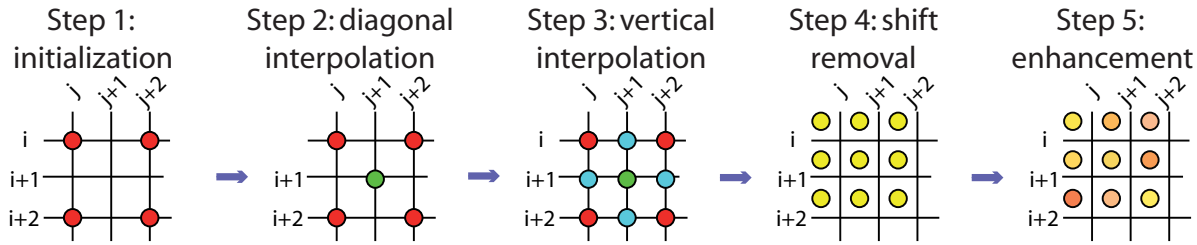


Figure 5.3: DFI steps sequence

ones available at a larger scale.

5.1.2 DFI algorithm in 5 steps: an overview

DFI interpolation is decomposed into 5 steps, as presented by figure 5.3. The initialization (step 1) creates an enlarged grid from low resolution pixels. Steps 2 and 3 fill the missing pixels by guessing neighbors' difference respectively in a diagonal and in an horizontal-vertical manner. Steps 4 and 5 can be considered as additional steps: step 4 corrects a $1/2$ pixel shift introduced by step 1 while step 5 is a quality enhancement step. The method is conceived to double the image size from $((N \times M)$ to $(2N \times 2M))$, but it can be iteratively repeated to reach any 2^n enlargement.

In the following, I denotes the low resolution image of size $N \times M$, $\tilde{I}$ stands for the interpolated image of size $2N \times 2M$, and $I(i, j)$ is a pixel located at the coordinates (i, j) in an image I.

5.1.3 Step 1 (Initialization step): pixel copy in an enlarged grid

DFI method classically enlarges images by first copying low resolution pixels in a larger grid of size $2N \times 2M$ as illustrated in figure 5.4. Therefore, $\frac{3}{4}$ of pixels are missing at this step.

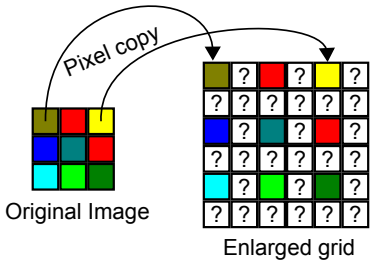


Figure 5.4: Step 1 - Pixel copy over an enlarged grid

5.1.4 Step 2: diagonal interpolation

Interpolation step 2 takes as input the enlarged grid created at the step 1 (red pixels in figure 5.3). Step 2 interpolates missing pixels that have four known neighbors on their diagonals. The pixels processed by the step 2 are visualized in green in figure 5.3.

In order to estimate the pixel value, the process computes the differences between the missing pixel and its fours known neighbors. The differences are obtained by following the geometry similarity principle presented in figure 5.2 by computing differences in the same direction but at a larger scale. This diagonal interpolation process is illustrated in figure 5.5.

As an example, the difference Δa between the missing pixel $I(2i+1, 2j+1)$ (pixels 2 in figure 5.5) and its top left neighbor is approximated by the mean value $\Delta a = (\Delta a1 + \Delta a2 + \Delta a3 + \Delta a4)/4$, where the set $\{\Delta ai\}$ represents the local differences obtained within the four quadrants. Once the missing pixel is estimated from its 4 neighbors, the resulting pixel value is bounded by the values of these 4 neighbors. The corresponding complete algorithm can be found in [182].

5.1.5 Step 3: vertical - horizontal interpolation

This second step benefits from the results of the first step. This step can be seen as the same as step 2 but rotated by 45 degrees. Step 3 interpolates missing pixels that have four neighbors in vertical and horizontal directions. The pixels processed by the step 3 are represented in blue in figure 5.3.

Step 3 algorithm is very similar to the previous one except that the immediate neighbors are in vertical and horizontal configuration. Identically, the missing pixel is interpolated by estimating its differences to its 4 neighbors. Once again, the differences $(\Delta a, \Delta b, \Delta c, \Delta d)$ are obtained by following the geometry similarity principle presented in figure 5.2 by computing differences in the same direction but at a larger scale (figure 5.6).

5.1.6 Step 4: 1/2 pixel shift

Step 1 introduces a $1/2$ pixel shift in the image as it does not spread evenly the low resolution pixel in the enlarged image. Ideally, the first low resolution pixel $I(1, 1)$ should be projected in the

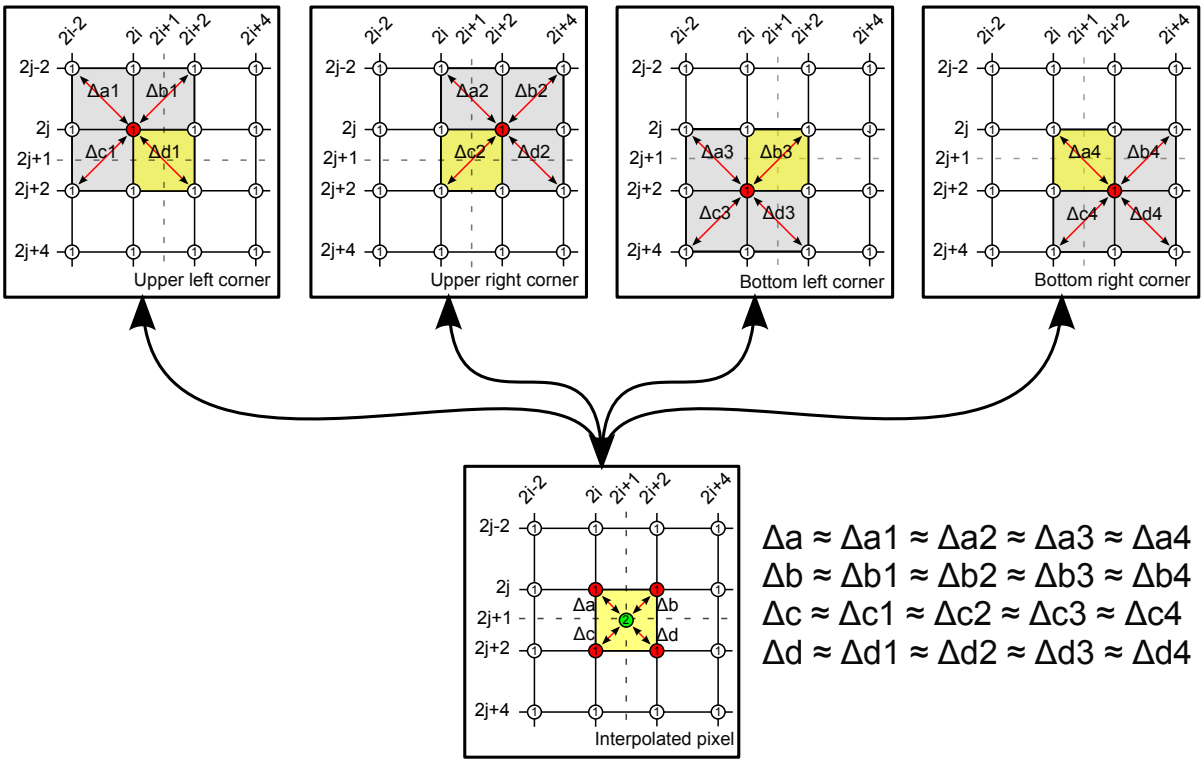


Figure 5.5: Step 2 - diagonal interpolation

interpolated image on coordinates $\tilde{I}(1.5, 1.5)$.

The step 4 is dedicated to compensate this shift. This step is identical to the step 2 and is applied to the interpolated high resolution image so that to create new pixels between pixels at integer position from a diagonal neighborhood (like step 2). Another way of seeing this is that it creates new pixels on an imaginary grid of the same resolution but shifted of a $1/2$ pixel position. Then, this imaginary grid substitutes the original one compensating the $1/2$ pixel shift (figure 5.7).

Regarding WPSNR scores, the shift correction is required in order to avoid a shift between the reference image and the interpolated image. In all other cases, the $1/2$ pixel shift can be optional, especially when the real high resolution image is unknown and when the image borders are off interest.

5.1.7 Step 5: Quality enhancement - local mean correction

Image interpolation is an up-sampling operation, leading to enlargie the image size, while down-sampling is the inverse operation. Ideally an interpolation operation followed by a down-sampling operation must result in an identity operation. This last step 5 aims at verifying that up-sampling and down-sampling operations are symmetric. If not, the step 5 corrects the interpolated image

In practice, any interpolation methods can introduce a bias that may need to be corrected. The presented quality enhancement step is a generic method that can be applied to any interpolation method. The algorithm **??** is then implemented the simplest way possible: in particular, other advanced down-sample filters can be used so that to increase the overall image quality.

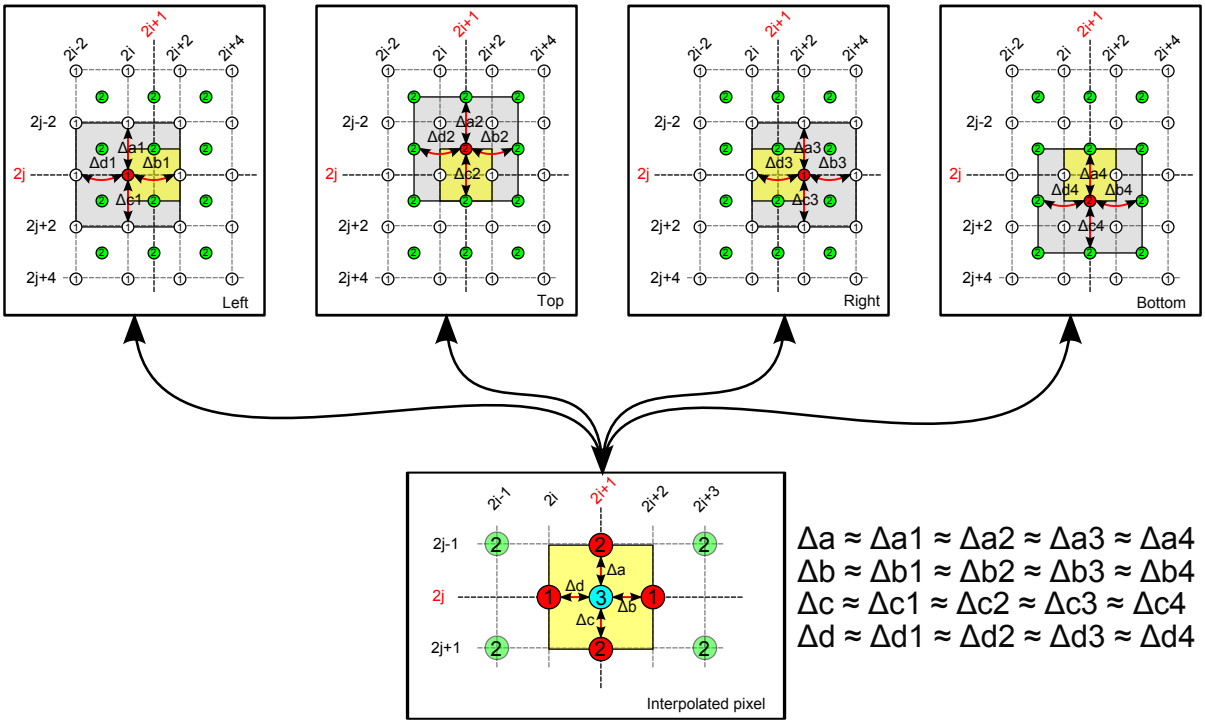


Figure 5.6: Step 3 - vertical / horizontal interpolation

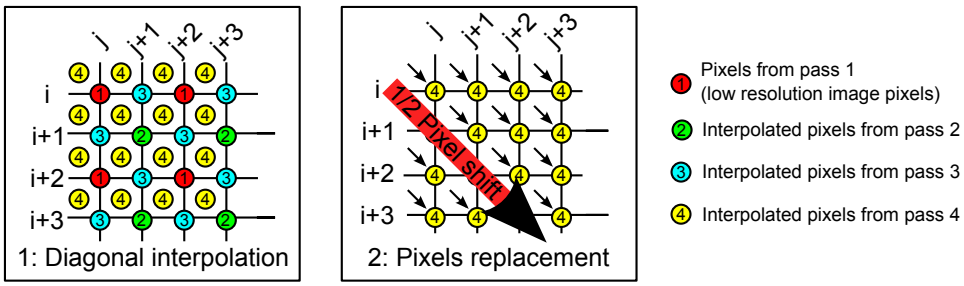


Figure 5.7: Step 4: diagonal interpolation for 1/2 pixel shifting

Local mean correction - Algorithm During interpolation, one pixel in the low resolution image I is projected over four pixels in $\tilde{I}$. The step 5 algorithm checks the projection in the opposite direction. It verifies that the interpolated pixels can be correctly projected back to the low resolution input image. Then, the correction algorithm checks if the mean of the four projected pixels in $\tilde{I}$ is equal to the corresponding pixel in I. If a correction is required, the four pixels are uniformly modified to get closer to the awaited mean.

Visually images appear sharper, however local mean correction also tends to add some aliasing on thin edges. The PSNR score of interpolated images are higher with the correction. The table 5.1 shows an increase of 0.84 dB over 25 images.

(a) Original image (b) Input image (c) Interpolated image WITHOUT quality enhancement (d) Interpolated image WITH quality enhancement

Figure 5.8: Quality enhancement step illustration, 8× enlargement

methods	mean WPSNR in dB over 25 images
Interpolation alone	32.08
Interpolation + local mean correction	32.92
Interpolation + smoothed local mean correction	32.87

Table 5.1: Local mean correction results on proposed interpolation method, average WPSNR scores on 25 images

Smoother local mean correction. To soften the aliasing effect, one solution is to apply a smoother correction. The correction must be especially smoothed on strong edges where the aliasing is more visible. The correction is thus not directly applied. In a first time, a raw correction map is computed containing pixel by pixel correction. In a second time, $\tilde{I}$ is corrected by a smoothed version of the correction map. By smoothing the correction map, the aliasing effect is strongly reduced together with keeping quality gain. Despite a reduced efficiency (-0.05dB) resulting images appear visually better with less aliasing (table 5.1). Figure 5.8 illustrates the quality enhancement step effects.

5.1.8 Border handling

On a general point of view, some methods do not handle image border well: either the borders are not processed or the output image is cropped and borders are simply removed. In order to handle the image border, the low resolution image is extended by border replication. This is a classical approach for image border filtering without modifying the filtering algorithm itself. In our case, a band of 8 pixels is added all around the image border. The band is created by symmetry with the pixel in the image. This image extension is illustrated in figure 5.9. Image extension by symmetry creates a complete neighborhood for border pixels and allows the DFI method to stay the same for all pixels.

5.1.9 Resulting images and objective quality

For evaluation purposes, we have produced images by applying all the 5 steps of the DFI method. In particular, the 1/2 pixel shift (step 4) and the quality enhancement (step 5) are used. Images are compared in terms of objective quality and in terms of subjective quality. For objective quality

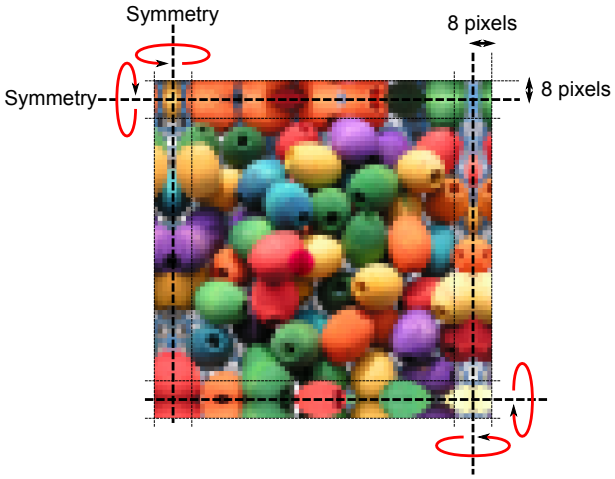


Figure 5.9: Image extension by symmetry for border handling

measures, 25 images are used. For subjective quality measures, a subset of 13 images are used. The complete testing procedure is described in [182].

5.1.9.1 Objective measure

Objective measures operate comparisons of our DFI method with a classical linear method of similar complexity. Table 5.2 shows a WPSNR gain for the DFI method of 0.5 dB when compared to the Lanczos method.

method	Bilinear*	Mitchell*	Lanczos*	**DFI method**
WPSNR	29.23 dB	31.10 dB	32.43 dB	**32.92 dB**

Table 5.2: Average WPSNR scores on 25 images of 512 by 512 pixels size. 2× enlargement. * : ImageMagick filter

5.1.9.2 Subjective evaluation

In order to evaluate the DFI method in a subjective manner, a subjective quality campaign has been performed on several images. The goal of such a campaign is to get an evaluation of the perceived visual quality of the DFI method and to figure how human subject feels about the DFI method comparatively to other methods.

The results, summarized in figure 5.10, consists of a subjective evaluation on 13 images, each image is interpolated with 8 different methods and rated by 14 human subjects. This type of assessment campaign is very time consuming and test subjects are voluntary.

The evaluation process is done as follows. For each image, the 8 different methods are presented together with the original image. Test subject rate freely each resulting image on a 0 to 10 scale. 10 being the highest quality. Images were interpolated from a size of 128 by 128 pixels up to a size of 512 by 512 pixels. The compared methods have a wide spectrum of complexity and are listed in the following sections.

On the low complexity side, 3 linear filter based methods, from the ImageMagick library [37], are used together with our DFI method, *i.e.* the nearest neighbor interpolation, bicubic and Lanczos interpolation solutions. Our DFI method has the same class of complexity than linear filters as presented in section 5.1.10. On the medium to high complexity side, 4 methods are used. ICBI [61] is a method with medium complexity. iNEDI [7] interpolation process is an improved version of [106] and is the most complex method of the set. The Pseudo LAD [165] of the "SAR Image Processor" software is used as it is often cited in the field and is well considered. PhotoZoom 3 [113] is a commercial enlargement software for photographers that has been also used in this test.

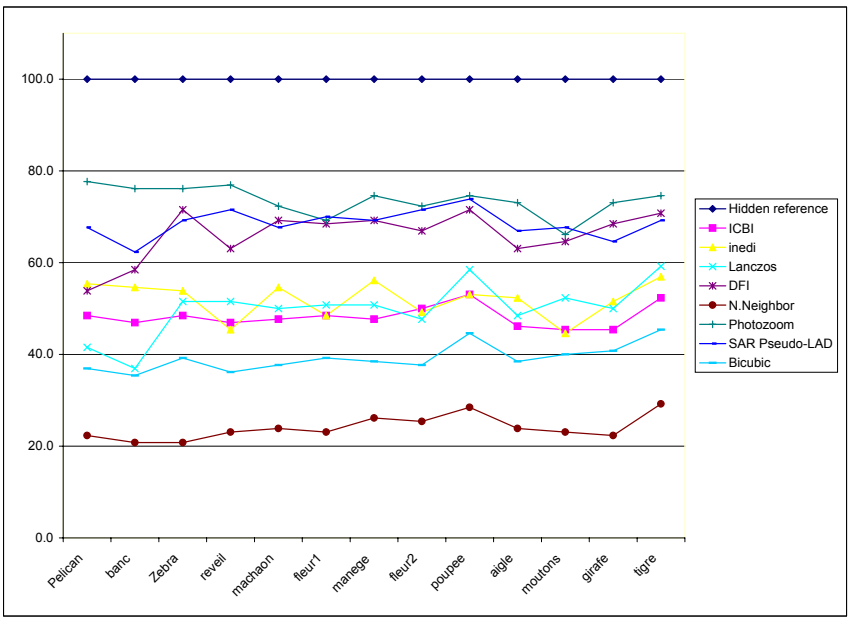


Figure 5.10: Subjective evaluation

From the figure 5.10, a group of 3 methods is clearly above the others and the DFI method is among them. Apart from the "Pelican" image, the DFI method gives rather close subjective scores to much more complex methods (PhotoZoom [113] and Pseudo LAD [165]). A second group of methods including ICBI, iNEDI, and Lanczos methods gives medium results with a clear quality gap from the first group of methods. In [61], the subjective evaluation of the ICBI method gives similar results to the iNEDI method. Our evaluation thus corroborates this equivalence. Lastly, bicubic and nearest neighbor methods give the worst results.

5.1.10 Complexity analysis and parallel implementation

This section presents a deep complexity analysis of the DFI method. In addition, the computational times of all the methods assessed in the previous subjective evaluation are presented.

Interpolation Algorithm analysis. The main part of our interpolation algorithm, *i.e.* steps 1 up to 3, operates only on integer data and does not use any multiplication. Operations are then restricted to bitshifts comparisons, and additions. Table 5.3 shows the number of each operation for each of the 5 steps. In addition, table 5.3 shows the percentage of pixels that are processed by each step.

Number of operations per pixel	+,-	bitshifts	comparisons	% of processed pixels
Step 1	0	0	0	25%
Step 2	21	5	4	25%
Step 3	21	5	4	50%
Step 4	21	5	4	100%
Step 5	10	2	4	100%
Mean operation number per pixel	46.75	10.75	11	100%

Table 5.3: Number of operations per pixel in the high resolution image for the proposed interpolation algorithm

Compared to other methods, the proposed DFI method exhibits a very low computational complexity. Moreover, the implementation uses integer data leading to faster computation time. For comparison purposes, the NEDI method [106] claims to perform 1300 multiplications per pixels.

Speed comparison. In order to illustrate its low complexity, our DFI method is compared to ImageMagick [37] interpolation functions and to the methods used in the subjective evaluation (section 5.1.9.2). Table 5.4 shows that the speed relative to the presented interpolation method is comparable to kernel based methods. This test shows that it is slightly faster that default ImageMagick filters but the DFI method achieves better visual results and lower memory consumption.

Interpolation algorithm (+ opening and saving)	time for 250 iteration	Peak memory
ImageMagick Lanczos filter	56.73 s	19404 ko
ImageMagick Mitchell filter	50.297 s	19384 ko
Proposed interpolation method	**49.50** s	**10504 ko**

Table 5.4: Benchmark of interpolation methods, enlargement by 2 of a 512 by 512 image

For the subjective evaluation (section 5.1.9.2), the DFI method has comparable quality results with Photozoom [113] and pseudo LAD [165] methods. However, the DFI method is at least 50 times faster (see table 5.5). As a consequence, the DFI method has the best tradeoff between quality and complexity of all the tested methods.

method	execution time	implementation
DFI method	$\approx$ 0.06s	compiled + multithreading
ImageMagick filters [37]	$\approx$ 0.06s	compiled + multithreading
ICBI [61]	$\approx$ 1.2s according to [61]	compiled
iNEDI [7]	>1000s	Matlab
pseudo LAD [165]	3 to 4s	compiled + multithreading
photozoom [113]	3 to 4s	compiled + multithreading

Table 5.5: Execution times of the methods used in the subjective evaluation: 128 by 128 pixels image enlarged to a 512 by 512 pixels image

5.1.10.1 Speed enhancement solutions

The goal of the presented DFI is to achieve the higher quality with the minimum complexity. If the DFI has proved to be a fast method, its speed can be enhanced at a cost of a slight lost of quality. Different solutions have been envisaged.

First a speed enhancement solution taking advantage of the YCbCr color space characteristics can be designed. Since each channel of the YCbCr does not have the same visual importance the interpolation process should be different from the Y channel to the others. Indeed, the Y channel has to be processed with more care than the chrominance channels that can be processed by simpler and faster algorithm. In this case, the Y channel can be processed by our interpolation algorithm while the chrominance channels are processed by a simple bilinear filtering. A WPSNR decrease of 0.278dB is observed when compared to the version where all color channels are processed by the DFI method. However, the time gain between the two approaches is more than 50% while the loss of WPSNR is only 1%.

Secondly, a speed enhancement solution relies on an image activity classification. Indeed, for interpolating algorithms, the most critical results are located on edges and high energy or textured regions. On those areas, the interpolation method quality is of the most importance to get satisfying visual results. On the opposite side, smoother areas (low energy regions) of the image can be processed with less care for the same visual final result. Our idea is to choose the proper interpolation method for each pixel according to the pixel energy class. Since low energy pixels can be easily interpolated with a simpler and faster method, the overall complexity can be reduced compare to the case where all pixels are interpolated with our interpolation method. As an example, within the LAR framework, the quadtree structure gives us clues on the nature of the pixels. Typically, 2 by 2 pixels blocks represent areas of highest activity, such as texture or contour, being more difficult to interpolate. The proposed solution is then to process with the DFI method all the pixels represented by 2 by 2 blocks in the quadtree, while the rest of the image will be processed by a faster bilinear filtering. In addition, surrounding pixel of 2 by 2 blocks will also be processed by the DFI method to avoid frontier artifacts. Results show a negligible loss of WPSNR (0.0017 dB) while the speed gain is about 12%. The acceleration factor strongly depends on the ratio of classified pixels linked to the quadtree threshold T_H and the nature of the image. On the 25 tested images, the speed gain is ranging from 0% up to 80% with similar WPSNR losses.

In any cases, the Hybrid DFI algorithm leads to speed gain with a minimal quality loss. Classification is here based on a quadtree decomposition of the image. However, any other classifier can be used, such as contour extraction in the low resolution image. Hybrid interpolation approach is therefore a general principle that can be applied on a wide application field.

For a maximal efficiency, both acceleration solutions can be combined. A speed increment of 59% with a slight quality loss of 0.277db is then observed. Figure 5.11 illustrates the visual effect of the hybrid solution. In particular, visually, the figures 5.11a and 5.11b are very similar.

Considering our speed performances, the DFI method is thus fast enough to be used for real time video interpolation. For example, it could be used to interpolate from SD to HD resolution. However, the DFI method is limited to enlargement ratio of 2^n and the ratio between SD to HD is 1.5. A solution exists to overcome this issue: it consists of adapting the DFI algorithm during the computation of missing pixels, as described in [32], by using a rectangular neighborhood of 1.5 ratio, instead of using a square neighborhood.

(a) Interpolated image, no acceleration

(b) Combined approach: YCbCr color space + Hybrid DFI

(c) Pixel classification for Hybrid DFI; white = DFI, black = bilinear interpolation

Figure 5.11: Combined approach: YCbCr color space + Hybrid DFI illustration, 2× enlargement

5.2 Region segmentation from quadtree structure

While its coding cost remains minimal, a quadtree contains a certain level of pseudo-semantic. The purpose of the presented algorithm is to extract the remaining semantic information from only the quadtree structure knowledge into regions. The idea is here to validate the assumption that the quadtree itself can be a clue of the image content. As no other information than the quadtree structure itself is required, the resulting segmentation map is sub-optimal but the overall complexity of the algorithm is very low. This solution can be then seen as a preprocess for further and finer segmentation processes.

The segmentation method relies then on two strong hypothesis. First, we assume that object contours are represented by smaller blocks. Secondly, we assume that the inner part of the objects is represented by bigger blocks in the quadtree that localize the object center.

Based upon these assumptions, the idea behind the method is to used a region growing approach, where the seeds are located on the bigger blocks. The regions will then grow by accumulating available surrounding blocks of decreasing sizes. Regions should then grow in objects from the inside out, and should automatically stop at object contour.

This original segmentation method has been designed by Clément Strauss during its PhD work.

5.2.1 Notations

In order to describe the segmentation algorithm, let us introduce some notations.

- Let QT be the the dyadic quadtree partition composed of P square blocks b_i, for $i \in \{1..P\}$ and where each block b_i has a surface of $2^S \times 2^S$ pixels, $S \in \{1..MaxSize\}$,
- The region segmentation method transforms the initial partition Δ^0 being the quadtree partitioning QT into a partition Δ^k $(0 < k < P)$ through k merging steps. Consequently a Δ^k partition is then composed of $(P-k)$ non overlapping regions R_i^k. Each region R_i^0 in Δ^0 then corresponds to a block b_i.
- Let $surf(R_i^k)$ be the surface in pixels of R_i^k,

- Let $|(R_i^k)|$ be the cardinal of R_i^k,
- Let A_i^k be the set of adjacent regions of $R_i^k \in \Delta^k$.

In order to easily describe the algorithm, the term "block" is used but actually refers to a "single block region".

5.2.2 Region segmentation algorithm

As previously mentioned, the segmentation algorithm 5.1) is initialized by the quadtree partitioning Δ^0. The algorithm can be then decomposed into two iterative processes, namely {Seeds creation} and {Region growing}.

First, the {Seed creation} process is executed. Once a first block of surface $CurrentSurf$ is added to the seed of a region, all adjacent blocks of surface $CurrentSurf$ are recursively added to the seed. During this step a set of seeds is created. Figure 5.12c illustrates the result of the creation of a set of seeds.

During the second process {Region growing}, all regions, issued from a seed, are simultaneously grown together. Regions are grown by merging their immediately adjacent blocks. Figure 5.12d illustrates the growth of seeds created in the previous step (figure 5.12c).

Neighbor blocks are merged if their surface is above the $CurrentSurf$ threshold. Once a first region has grown by merging its immediate neighborhood, the next region is grown until all regions are processed. This last process is repeated a given number of times namely $iter(CurrentSurf)$ times. $iter(CurrentSurf)$ is here manually set up considering the awaited results and the context.

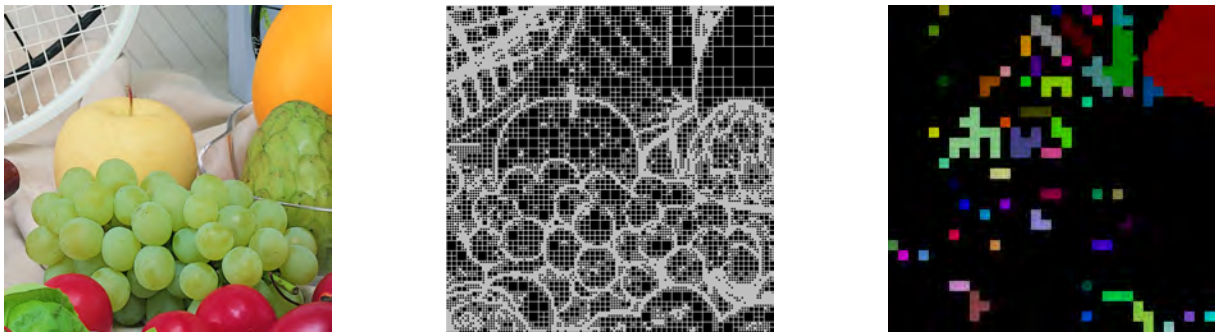
Making regions grow one by one leads to a simultaneous region growing process. In our setup the number of growing repetitions increases when the block size threshold $CurrentSurf$ decreases. The more growing repetitions are performed, the more blocks are merged thus decreasing the total number of regions.

The two processes are re-iterated with a reduced value of $CurrentSurf$. New seeds are created, leading to new regions, then all regions are grown by merging surrounding blocks that match the $CurrentSurf$ threshold. Figure 5.12 illustrates few steps of the region segmentation algorithm.

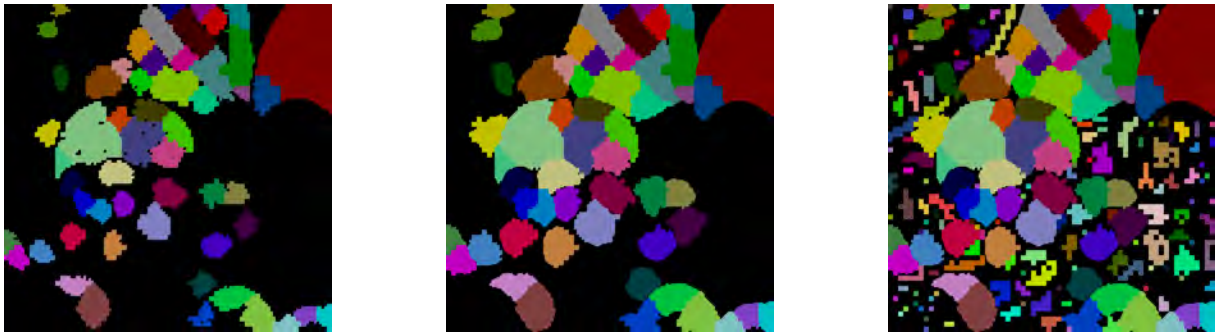
The algorithm does not consider the pixel value and only uses the quadtree structure without additional data and without supervision. In this sense the algorithm performs a kind of blind segmentation.

5.2.3 Performances analyses

Segmentations may seem to exhibit a rather high number of regions as shown in figure 5.12. Keeping a rather high number of regions is justified by the limited content of the input data (quadtree) as shown in figure 5.12f. Moreover a large number of regions also mean smaller regions thus avoiding strong segmentation errors. Indeed if a region grows too much it is more likely to cross over a semantic contour where it should have stopped. It also enables a more accurate visual representation. The data given by the lone quadtree is so light in terms of information that it is difficult to extract big regions covering a whole semantic object. Therefore a semantic object will be more likely split into several regions. This solution has been chosen in order to avoid the case where a single region overlaps two different objects.



(a) Original image (b) Quadtree (c) New seeds creation



(d) Region growing step 1 (e) Region growing step 2 (f) New seeds creation

Figure 5.12: Few steps of the segmentation method in false colors

Algorithm 5.1: Complete region segmentation algorithm

```
/* Initializations                                                    */
k = 0;
Δ^k = Qt;
CurrentSurf = 2^MaxSize × 2^MaxSize;
repeat
    /* Seeds creation                                                 */
    while ∃ R_i^k |( surf(R_i^k) = CurrentSurf and |R_i^k| = 1) do
        while ∃ R_j^k ∈ A_i^k and surf(R_j^k) = CurrentSurf and |R_j^k| = 1 do
            Merge R_j^k and R_i^k into Δ^{k+1};
            k = k + 1;
            Update A_i^k;
        end
    end
    CurrentSurf = ⌊CurrentSurf/4⌋;
    /* Region growing                                                 */
    while ∃ R_i^k | surf(R_i^k) = CurrentSurf do
        for iter = 1 to iter(CurrentSurf) do
            Let A' = {R_j^0 | (R_j^0 ∈ A_i^k and surf(R_j^0) = CurrentSurf and |R_j^0| = 1)} ;
            Let N = |A'|;
            Merge R_i^k and A' into Δ^{k+N};
            k = k + N;
            Update A_i^k;
        end
    end
until CurrentSurf = 0;
```

5.2.3.1 Evaluation of the segmentation

Evaluating a segmentation is a hard task [92]. Segmentation evaluation can be performed with help of standard databases. Another way of evaluating a method can be obtained by evaluating the application enabled by the region representation. It can give a idea of how good is the segmentation for a given task.

The Berkeley image segmentation database [124] measures segmentation processes by comparing region contours with human segmentation. This database is widely used for segmentation evaluation purposes. The Berkeley database contains both test images and training images with human ground truth. In addition a benchmark method is given.

Considering this image database, it cannot be directly used in our case since our pseudo semantic level is not high enough. However the comparison methodology can still be exploited.

As for the benchmark, our segmentation process produces too much regions comparing to what human subjects do. Therefore using the Berkeley benchmark method is not a valid option [92]. In addition our segmentation is working from a reduced data set (quadtree) that already reduces the semantic. Similar method in that way does not exist to our knowledge and comparing our method

with other methods with full access to the image would be meaningless.

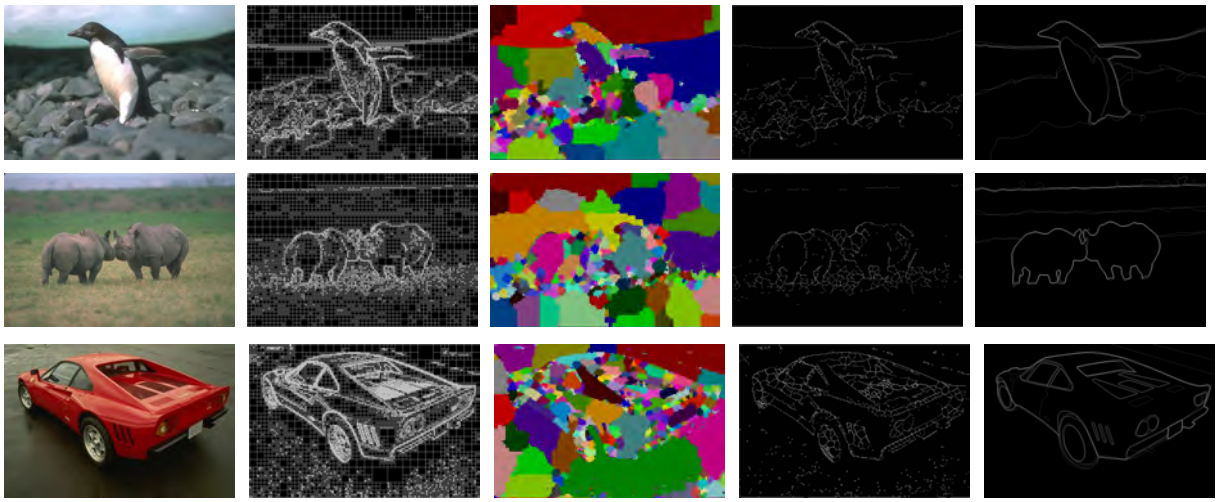
Even if the Berkeley segmentation benchmark cannot be used as a whole, our method can, nevertheless, be visually compared either to the proposed hand drawn ground truth or relatively to the original image.

Figure 5.14 and 5.13 present segmentation results obtained on 3 images of the Berkeley database [124]. Each image is presented with the corresponding quadtree, the segmentation map in false colors, and pseudo region boundary extraction

The region pseudo contour image shows region boundaries only where the quadtree exhibits 4 by 4 or 2 by 2 pixel blocks. Region boundaries are shown in white when they correspond to 2 by 2 blocks and in grey for 4 by 4 blocks. Displaying only contours on smaller blocks enables to highlight one of our hypothesis on the quadtree block size semantic. Indeed we assume that smaller blocks semantically correspond to object contours.

Figures 5.13 and 5.14 show that our segmentation process manages to extract from the quadtree some object contours. This demonstrates that the presented method manages to reach a certain pseudo semantic level and validates both our region growing approach and our hypothesis about the quadtree block semantics.

As previously mentioned, a rating of our segmentation is difficult and comparing with other approaches has no real signification according to our specificities. However both segmentation maps and pseudo region contours partially exhibit the pseudo semantic contained in the quadtree and partially retrieve the semantic contained in the original image.

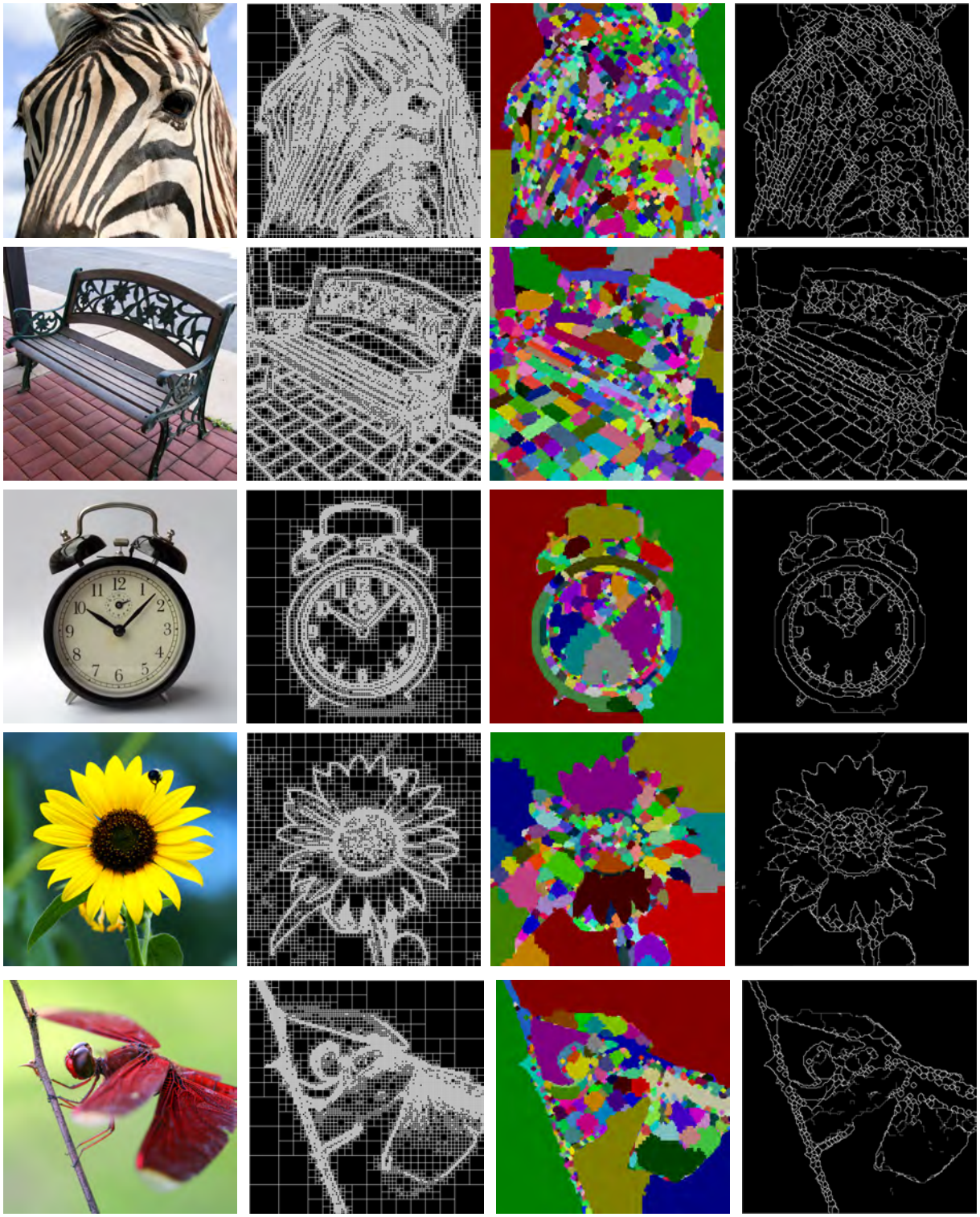


(a) Original image (b) Quadtree (c) Segmentation (d) Method pseudo contours (e) Ground truth

Figure 5.13: Segmentation method evaluation, Berkeley database

5.2.3.2 Complexity analysis

Region segmentation process is composed of two major steps: quadtree decomposition and region merging. Quadtree decomposition has a low complexity and does not use any multiplication, division, or floating point operations. It uses mainly comparisons to compute the local homogeneity.



(a) Original image (b) Quadtree input (c) Segmentation in false colors (d) Method pseudo contours

Figure 5.14: Segmentation method evaluation

Region merging itself also has a low complexity. First it uses a reduced amount of data since the quadtree strongly sums up the image content. Secondly the region merging depends only on comparison operators and no multiplication is involved. In addition its complexity is proportional to the image size. In the followings, execution times are given for a C implementation on an Intel core 2 duo at 2.5 GHz used with a single core. The quadtree split process takes about 10ms and the segmentation process typically takes from 30ms up to 60ms for a 512 by 512 pixels color image. These could be optimized by tuning the implementation.

5.2.3.3 Potential applications

The image segmentation method is based on a strongly reduced dataset. Segmentation is not directly performed in the pixel domain. Instead, the method uses an empty quadtree structure, assuming that this structure matches the semantic content of the image. Several algorithms and codecs embed such quadtree structures from where the method could operate. Despite the low semantic content of the input, the segmentation method manages to retrieve some semantic contained in the original image and approaches the ground truth results. The input data being so light, our method has a very low complexity even when considering the quadtree partitioning. The results show that both our hypothesis about the semantically compliant quadtree structure and the region growing approach make sense. The seed selection efficiently selects object center, and the region growing based on block sizes manages to recover some original semantic.

In addition the quality of our results show that a quadtree partitioning enables an efficient speed up for image segmentation by simplifying the information without destroying important semantic parts. This quadtree partitioning approach should be considered as a speed up approach for other methods.

One advantage of our method is to enable a basic segmentation without needing color information. As an example, in a video or image coding scheme, this could enable a semantically adaptive processing prior to pixel value availability or knowledge. However, as this method does not use color information, adding such information should naturally improve the segmentation representation by suppressing some uncertainties during the region growing algorithm, hence improving the segmentation quality, at the expense of complexity. It also could be seen as a preprocess for advanced segmentation solutions so that to accelerate the overall process.

5.3 Multiresolution segmentation

In parallel to the previous low complex segmentation solution, a scalable segmentation algorithm called JHMS (Joint Hierarchical and Multiresolution Segmentation) characterized by region-based hierarchy and resolution scalability, has been proposed by Rafiq Sekkal during his PhD work. Most of the state-of-the-art algorithms either apply a multiresolution segmentation [93, 197, 181] or a hierarchical segmentation [94, 127]. The proposed approach combines both multiresolution and hierarchical segmentation processes. Indeed, the image is considered as a set of images at different levels of resolution, where at each level a hierarchical segmentation is performed.

Multiresolution implies that a segmentation of a given level is reused in further segmentation processes operated at next levels so that to insure contour consistency between different resolutions. Each level of resolution provides a Region Adjacency Graph (RAG) that describes the neighborhood

relationships between regions within a given level of the multiresolution representation. Region label consistency is preserved thanks to a dedicated projection algorithm based on inter-level relationships.

Semantic or pseudo-semantic region extraction takes then advantage of multiresolution representation in terms of region definition at the expense of the computational complexity. Then an effective tradeoff should be found so that to obtain both content compliant representation and a fast segmentation solution.

This solution remains generic as the innovative part relies on the RAG description. This framework can be then coupled to any hierarchical segmentation system, such as

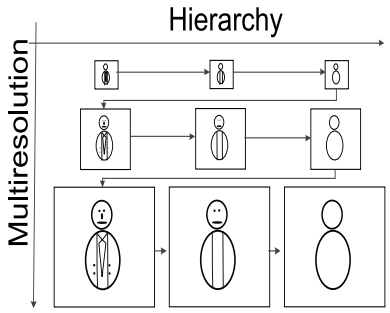


Figure 5.15: Multiresolution and hierarchical segmentation representations

5.3.0.4 General principles

The JHMS technique relies on an iterative process of the pyramid until reaching the desired level of resolution $lmin$. As described in figure 5.16, an initialization process from the low resolution image is first required.

Let RAG^l be the Region Adjacency Graph at resolution level l, with $l \in \{lmax, \ldots, lmin\}$. Regions of the initial RAG^{lmax} correspond to the blocks of the lowest resolution image. Then, a hierarchical segmentation modifies the RAG^{lmax} by including hierarchical description.

Iteratively, for each level l of the pyramid, the current RAG^l is obtained by projecting the RAG^{l+1} onto the RAG^l resolution. RAG^l is then updated by taking into account both the projected RAG^{l+1} and changes within neighborhood relationships. Finally, a hierarchical segmentation is performed onto the current RAG^l.

5.3.0.5 Multiresolution RAG

Figure 5.17 depicts the different steps of the multiresolution RAG algorithm. To project RAG^{l+1} onto RAG^l across resolution, the algorithm projects the regions labels by using the quadtree partition. Indeed it is used as a reference to decide which blocks are kept unchanged and which ones are split. Thus regions composed of at least one unchanged block are considered as fixed regions, and are directly projected in the current level. Labels of fixed regions are maintained for the next segmentation step. Inheriting labels ensures the label consistency of the segmentation. Unchanged blocks correspond to leaves of the quadtree. Thus if an unchanged block is detected at l^{th} level, it means that 2^{l+1} pixels in the full resolution will be discarded from the computation. Consequently, the computation is reduced thanks to quadtree partitioning.

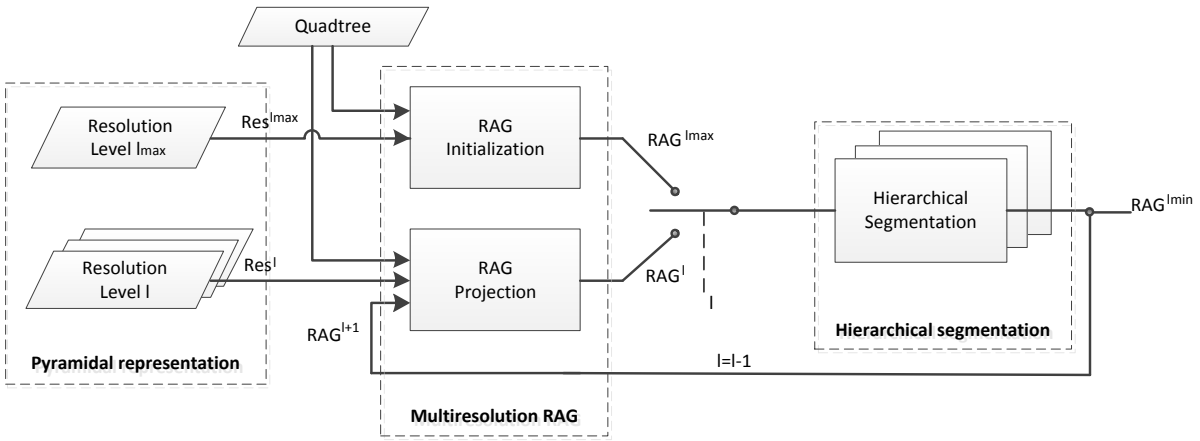


Figure 5.16: JHMS general scheme

Region relationships between two successive levels are shown in figure 5.17. Two kinds of region relationships are obtained: regions created in RAG^l with blocks from the same region in the RAG^{l+1} (Figure 5.17.a) or regions with blocks that belong to different regions in RAG^{l+1} (Figure 5.17.b).

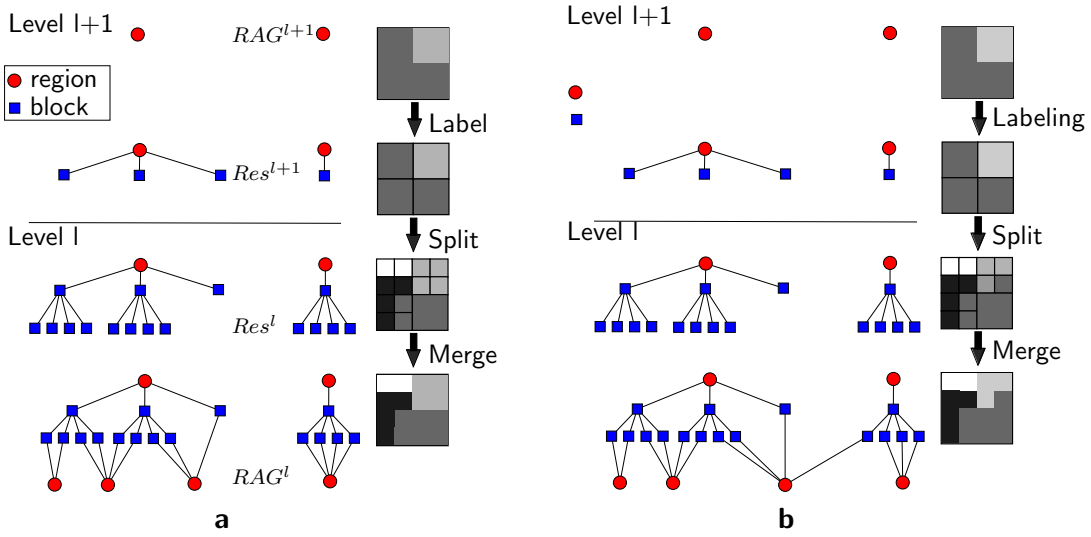


Figure 5.17: Inter-level regions relationships. Four regions in Res^l, in a) regions are composed with blocks of the same region parent. However in b) there is one region composed from blocks belonging to the two regions parents

5.3.0.6 Hierarchical segmentation

An extension to hierarchical representation at each level of multiresolution can be designed to overcome the natural resulting over-segmentation by selecting the segmentation granularity. This global solution called Joint Hierarchical and Multiresolution Segmentation (JHMS) provides a highly scalable region representation.

5.3.1 Experiments and results

The proposed algorithm combines two criteria. The first criterion is the difference between the mean values of two adjacent regions. The second criterion computes the gradient between blocks along the shared contour between two regions. For our experiments, we used the Locally Adaptive Resolution structure to build the quadtree and the hierarchical segmentation function (see section 2.2.3). Although a learning step is recommended to find the optimal set of parameters, they have been empirically tuned relying on experiments when trying to get effective results.

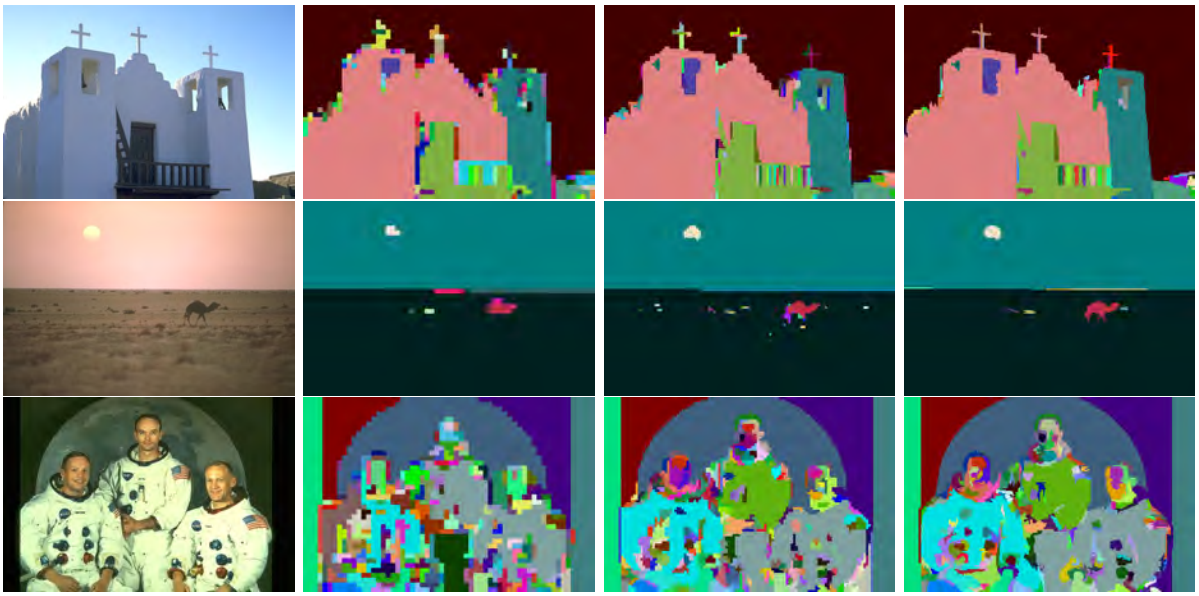


Figure 5.18: Scalable segmentation results

5.3.1.1 Visual results

In Figure 5.18, resulting segmented images are shown. Regions are here presented in false color. From the left to the right, original image is first presented, then the corresponding oversampled image of labels at different resolution levels (*i.e.* $3, 2, 1$). As can be observed, results at this resolution are satisfactory in terms of representation. In addition, the consistency of region labels is well preserved across resolutions, so that object tracking can be envisaged throughout the multiresolution. Furthermore, color gradated regions such as the sky of the first image are well detected as a single region thanks to the local gradient feature.

5.3.1.2 Objective quality of segmentation

In order to compare the proposed algorithm with the literature, obtained segmentation maps are tested against the Berkeley benchmark[124] . This benchmark is usually used for comparing contour detector algorithms. Contour maps are thus compared with contour maps designed by human beings which are then considered as ground truth. In our experiment the BSDS300 dataset is used.

The proposed algorithm is strictly based on local gradient and mean merging criteria. To provide a fair comparison, only color-based contour detectors, from the Berkeley benchmark, that share the same features have been compared. The algorithm results have been then compared with ground truth images. As shown in table 5.6, JHMS obtained F=0.60 score. As for it, GPB [117] (*Global Probability of Boundary*) combines the use of local information derived from brightness, color, and texture signals to produce a contour detector. It provides the best performance on the benchmark with F=0.70. CG[125] (Color Gradient), another algorithm based on same features as the JHMS provides a score of F=0.57.

JHMS provides a pseudo semantic segmentation and is not able to reach object level granularity by itself. Typically, the difference of score with GPB is due to the fact that additional texture information are used in the GPB segmentation. Texture improves the segmentation results, however, extracting texture features makes the algorithm more computationally complex.

Algorithms	Ground Truth	GPB	CG	JHMS
Scores Average	0.79	0.70	0.57	0.59

Table 5.6: Quantitative scores on Berkeley database BSDS30

In figure 5.19, different contour maps are presented from GPB, CG and JHMS techniques. Most of contours in JHMS are similar to those found in the ground truth images. Neither global information nor texture features are used in JHMS. In consequence, any strong brightness or color variations within a single object lead to over-segmentation, thus penalizing the score of our algorithm. For example, to segment the tiger in the last row of figure 5.19, hand segmented images consider the whole tiger as a single region and one consistent contour. With the proposed algorithm, the tiger is detected as multiple adjacent regions corresponding to the stripes of the skin.

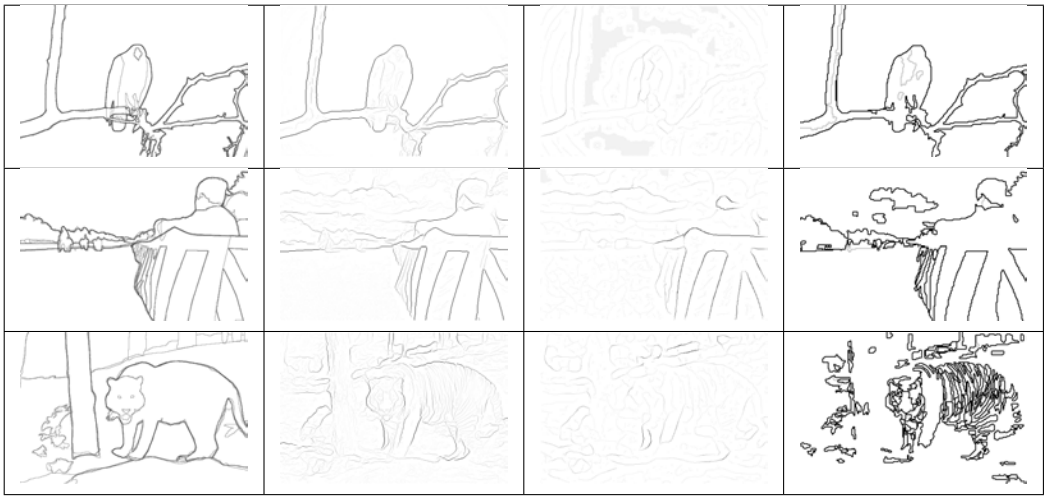


Figure 5.19: Image Boundaries, from the left: ground truth, Global Probability of Boundary GPB, Color Gradient (CG) and JHMS results

5.3.1.3 Multiresolution and quadtree partitioning influence on complexity and objective scores

In order to exhibit the influence of the quadtree partitioning and the influence of the multiresolution scalability on the computational complexity and objective scores, segmentations in different configurations have been performed on the BSDS300 dataset images. Results are here obtained using an Intel Core i7 @2.67GHz with a single thread. The average time of segmenting the 100 images of the dataset and mean scores of Berkeley benchmark have been measured.

First the influence of the quadtree partitioning is evaluated without multiresolution. Segmentations are directly performed on the full resolution image, either on pixel based images or on block based images following the quadtree partitioning. Table 5.7 shows that the quadtree partitioning enables a 5.5 times speed up compared to the pixel based images. In addition the objective score with quadtree partitioning is better. This can be explained by the fact that the quadtree helps by guiding the segmentation algorithm in the image by providing a first pseudo semantic description.

Secondly the multiresolution segmentations are compared with an increasing number of resolution levels. Table 5.7 depicts the results of our experiments with segmentations performed without multiresolution, and with up to 4 levels of embedded resolution levels. Objective scores remain almost identical when increasing the number of resolution levels and remain close to the performances of the segmentations without multiresolution. However, using multiresolution, scalability strongly impacts the complexity. Complexity comes from both the RAG projection from one resolution level to the next and the segmentation itself at one level of resolution. With more levels of resolution the segmentation at each resolution level is simplified, therefore, more levels tend to reduce the overall complexity. As for the RAG projection mechanism, its complexity is proportional to the number of blocks. From the half resolution up to the full resolution, the RAG projection handles much more small blocks than during other previous projection steps. Therefore, the last RAG projection onto the full resolution level explains for the most part the complexity of the method.

Multiresolution	none	none	2 levels	3 levels	4 levels
Quadtree partitioning	no	yes	yes	yes	yes
BSDS300 scores	0.58	**0.60**	0.575	0.577	0.578
Execution time (s)	1.100	**0.200**	1.429	1.355	1.300

Table 5.7: Multiresolution and quadtree partitioning influence on complexity and objective scores

5.4 Conclusion

In this chapter, we first describe an innovative and low complex interpolation solution. The DFI interpolation method was developed as a universal method and its only limitation is the 2^n enlargement ratio. Concerning the quality, subjective tests show that the DFI method is visually almost as good as state of the art method. However, the DFI complexity is much lower and comparable to linear methods and is at least 50 times faster than other tested methods with comparable or better quality. The DFI is highly parallel and has been implemented on multicore architectures with acceleration factor close to the number of cores. In addition, speed enhancement solution has been proposed, with help of color spaces and classification, enabling a speed increment of 59%, this with minimal

quality loss. Therefore, the DFI method has one of the best tradeoff between quality and complexity, and in this sense fulfill our objectives.

Then we focused on the quadtree structures, to extract some pseudo semantic content. In the Interleaved S+P codec case, this quadtree is very light to encode and does not contain much information, but is transmitted prior to anything else for a minimal cost. The ultimate goal of such semantic extraction is to enable content dependent processing without other knowledge than the quadtree structure itself.

In order to extract semantic directly from the quadtree structure, several methods have been developed that give different types of information. In order to be able to extract semantic at object level, a kind of blind region segmentation method has been developed that only uses the quadtree structure. This segmentation is labeled as "blind" because there is no color information about the image. This method does not give one region per object but ensures that an object is represented by a set of regions. Even if raw results in terms of region representation are not competitive with other segmentation method, mainly because of oversegmentation, yet we proved the ability of the quadtree structure to represent by itself the image content.

Finally, a RAG-based multiresolution segmentation has been designed. Coupled with any hierarchical segmentation process, the proposed solution is highly scalable while remaining efficient both in terms of computational complexity and in terms of region representation quality.

From these generic tools, dedicated algorithms could benefit from them. In particular, following chapters address two different potential application context, namely pseudo-semantic video codecs as well as vision robotics.

Chapter 6

Pseudo-semantic representation of videos: joint analysis and coding tools

Previous chapter was dedicated to generic analysis tools. In particular, quadtree-based segmentation processes, addressing different objectives, were presented. These methods can be coupled to any other method aiming at automatically analyzing the image content.

When considering video coding, the ability to locally describe the geometrical contents appears to be an essential feature so that to take advantage of the inherent redundancies. Indeed, most of the time, image sequences account for the evolution of a scene across time. As a consequence, each image can be interpreted as a deformed version of its predecessor in the sequence. For this reason, the largest part of the redundancies is located along the temporal axis and should be further studied.

The image content can be thus described by different levels of semantical meaning. If object-based extraction solutions integrate high level of semantical information, such as whole object shape, intermediate tools aim at obtaining a coherent representation in terms of texture, motion, color or any combination of this typical features. As they do not rely on any side information, these intermediate representations can be qualified as pseudo-semantical methods, as they tend to fit to the image content. In this context, disruptive approaches, relying on local cues analysis, can lead to pseudo-semantic video representation, able to exhibit the video content.

Three different pseudo-semantic representations of videos are proposed in this chapter. If they initially address coding issues, the resulting video representations can be used by itself to reach higher semantically consistent representation.

In section 6.1, a RAG-based video segmentation process is described. The main innovation relies in the ability to exhibit the persistence of the regions along a given sequence. This video segmentation can be used as a first step for object recognition, tracking, classification etc. It relies on a quadtree based segmentation representation, fully compliant with the ones presented in chapter 5.

Section 6.2 investigates the viability of an alternative representation that embeds features of both classical and disruptive approaches. Its goal is to exhibit the temporal persistence of the textural information, through a time-continuous description. However, it still relies on blocks, mostly responsible for the popularity of the classical approach. Instead of re-initializing the description at each frame, it is proposed to track the evolution of initial blocks taken from a reference image. A block, and its trajectory across time and space, is called a motion tube. An image sequence is then interpreted as a set of motion tubes.

In between the two previous solutions, an image analysis/synthesis framework, designed for video coding, combines the region segmentation and characterization tools together with texture synthesizers (section 6.3). Thus, it aims at improving the coding efficiency for textured regions in images and videos. The basic assumption is based on the human visual system properties, which prefers synthesized details to flat color, even if the output surface is not exactly the source texture. In this work, texture synthesis algorithms from the literature are adapted in order to fit the coding context. In particular, the idea is to fill textures which are not entirely transmitted.

6.1 Consistent spatio-temporal region representation

Defined as second generation coder, region-based approach tends to link digital systems and human perception. This type of approach provides advanced functionalities such that scalable video coding [130][122], Region of Interest (ROI) coding, video object tracking [28] and content manipulation [168]. Despite the benefits of region-based solution, actual standards MPEG4-AVC [70] and SVC [84][207] and even HEVC [208] are always based on traditional hybrid coding scheme. Principal obstacles to the content-based system evolution are

- the high amount of side data required to described region boundaries,
- the high complexity of segmentation algorithms.

As an example, the SESAME coder that achieves a rate-distortion (RD) optimization on a multiscale frame representation and explicitly encodes the resulting segmentation partition [123, 168]. Alternative solutions attempt to not transmit the region partition coding while only considering decoded information in the segmentation process. For example, [215] presents a symmetric-complexity coding system where both coder and decoder compute a motion segmentation in a video coding context. Unadapted for low bit-rate, this solution produces unacceptable region description especially upon contours.

During his partial PhD work (he did not wish to end and defend his thesis work), Erwan Flécher has designed a content-based color image sequence coding system. Principal improvement of this coder called LAR video relies on a temporal multi-scale representation of the video content that is not explicitly transmitted to the decoder. A hierarchical segmentation aims to efficiently compress both color components and motion information by region levels. Restricted to the Flat layer encoding, the LAR video can be seen as a natural extension of the LAR region based color image representation (see section 2.2.3) [55].

6.1.1 Region-based image sequence coding: framework

The proposed color image sequence coding system is based on Intra (I) and Predictive (P) frame compression where only the Flat layer is considered. On the one hand, luminance I-frames are efficiently compressed with the Flat Interleaved S+P coder (intra prediction mode). On the other hand, inter prediction mode that uses temporal redundancy between consecutive frames has been developed to compress luminance P-frames. LAR video coder takes advantage of on the multi-scale representation of the frame content that is presented both at coder and decoder (section 2.2.3). This allows the encoding of chromatic components and the definition of a region-based representation of the motion information (figure 6.1).

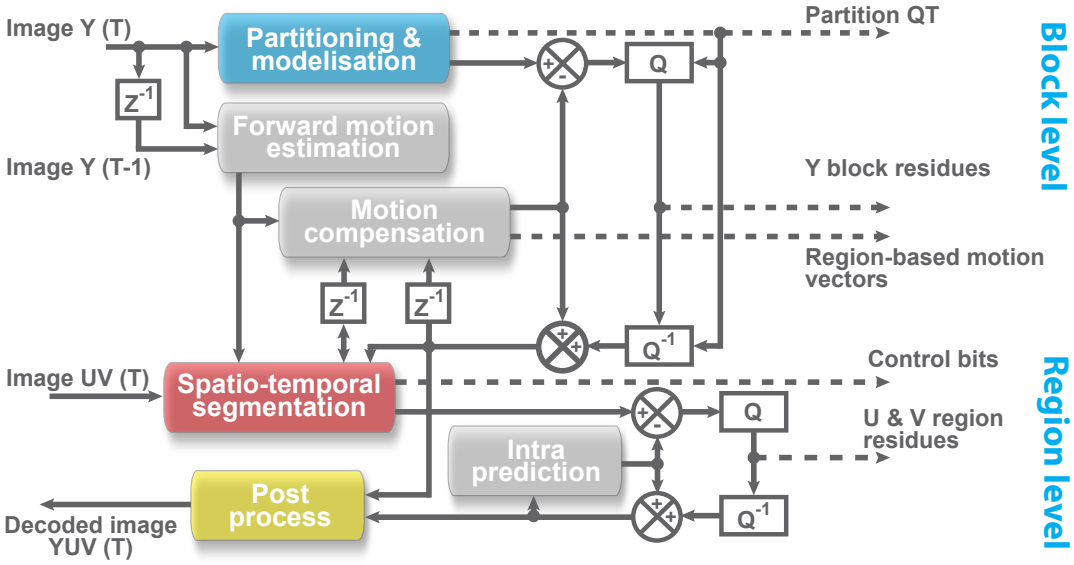


Figure 6.1: Region based video coding framework

The principle of the low complexity coder can be described with three sequential steps as schematically shown on figure 6.2:

Step 1 Luminance block image is compressed using inter/intra prediction modes and region-based motion compensation (figure 6.2.a),

Step 2 Considering Y-block image and chromatic control, multi-scale spatial segmentation for region-based chromatic components representation is performed (figure 6.2.b). It allows color image compression at region level and non-explicit definition of Partition Tree PT_s^N,

Step 3 Using motion feature, a hierarchical spatio-temporal segmentation is derivated of PT_s^N by split/merge process. It provides a compressed region-based description of motion between frames (figure 6.2.c).

As a consequence, for each frame, the bitstream includes block size information from the quadtree representation, quantized prediction errors of the Y-block image and region-based Cr/Cb data, prediction errors of region-based motion vectors and side information such as chromatic control bits and motion split/merge related data.

6.1.1.1 Luminance block image prediction and coding

In the LAR video coder, partition is directly computed on the current frame in order to provide the luminance block image previously described. Inter/intra prediction modes and quantization step are then used to reduce the bit-rate of the Y-block image. For P-frames, previous decoded frame is projected on the current quadtree partition using region-based motion vectors. Consequently, three types of blocks (defined in the current quadtree partition) result from the compensation step (see figure 6.3): blocks with only one projected value, blocks with multiple projected values (overlapped blocks) and blocks with zero projected value (uncovered blocks ot I frames). In two first cases, mean value is computed and is used as block prediction. High prediction errors (included overlapped

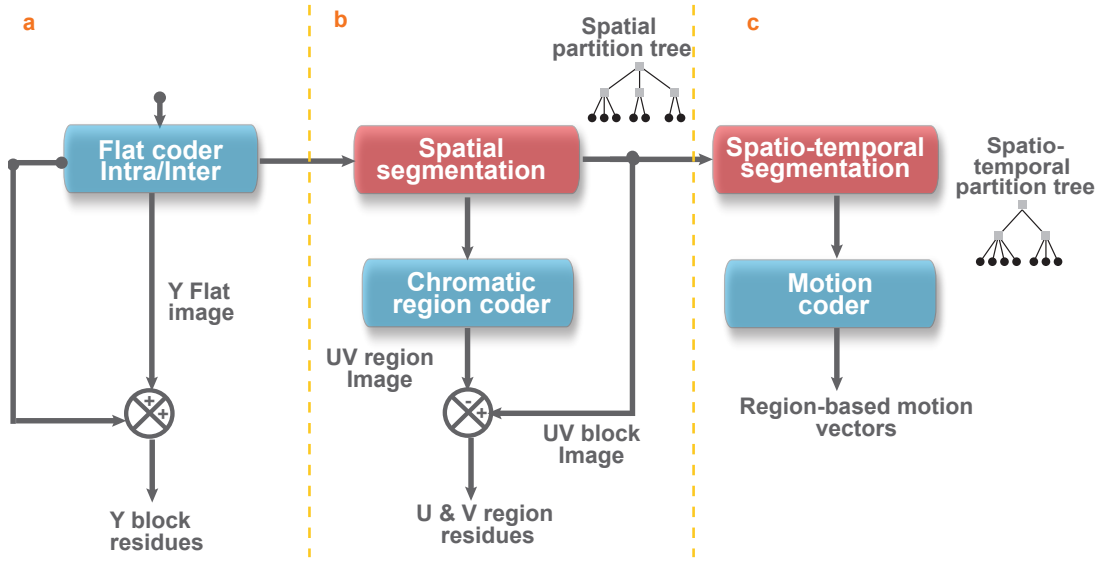


Figure 6.2: Simplified region based video coding framework

blocks) are classically located on moving object boundaries especially when region-based motion compensation is operated. DFD energy is thus mainly concentrated on small blocks which are efficiently compressed with a block-size adapted quantization. In last case, blocks are predicted in intra mode with the same edge-based predictor that is used for still image encoding.

6.1.1.2 Hierarchical spatio-temporal segmentation

Once Y-block image has been compressed and transmitted to the decoder, spatial segmentation that only considers information of the current frame is processed independently for I and P-frames. A region-based motion compensation has been considered for P-frame encoding. Indeed, a region-based motion estimation/compensation provides better performances that classical block-based approaches especially upon moving contours. As a consequence, spatio-temporal segmentation aims to provide a multi-scale representation in which defined regions respect spatial and temporal homogeneities. They thus naturally share common features such as gray level, color and motion. Figure 6.4 gives a basic description of the spatio-temporal segmentation and motion encoding. Considering that the Partition Tree PT_s^N is known at the decoder, the aim is to define a new Partition Tree PT_{st}^N based on the spatio-temporal segmentation. Sequential splitting and merging steps are realized in order to decompose regions with non-homogeneous motion and to group regions with spatial and temporal similarities. Information associated to motion-based splitting/merging is transmitted because the decoder does not know region-based motion parameters. Note that PT_{st}^N describes a research-space adapted to Rate/Distortion optimization thanks to the inherent indexed hierarchy [130].

To describe region-based motions, in our experiments, a translational motion model has been used. Low complex parameter computation and efficient parameter prediction have motivated the choice of this model. In order to take into account motion information, a well-known fast and variable block-size motion estimator called EPZS (Enhanced Predictive Zonal Search) [192] is initially used to compute the forward motion vector of each region (or block) of the partition. Motion estimation on supports with inconsistency size (classically 2×2 block-size) are not estimated but result from

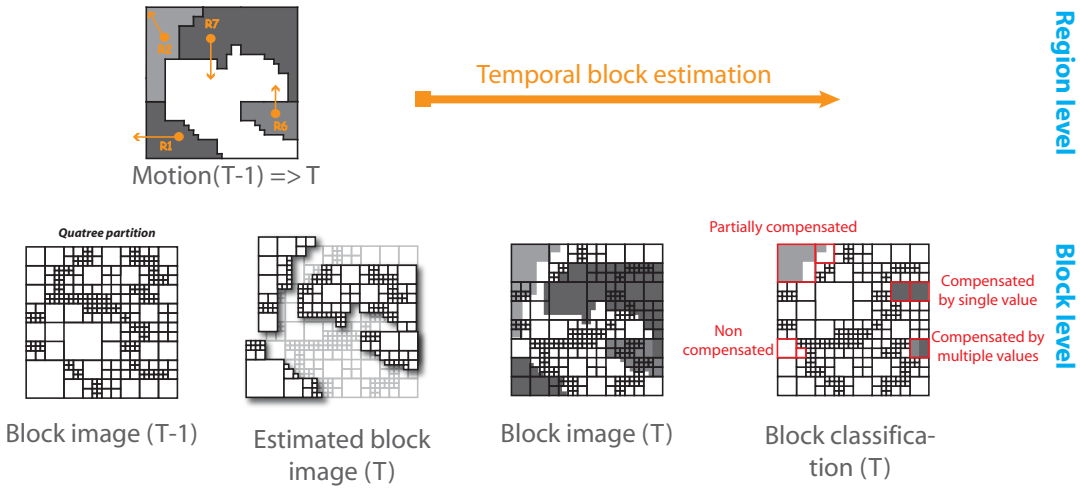


Figure 6.3: Motion compensation and quadtree partitioning

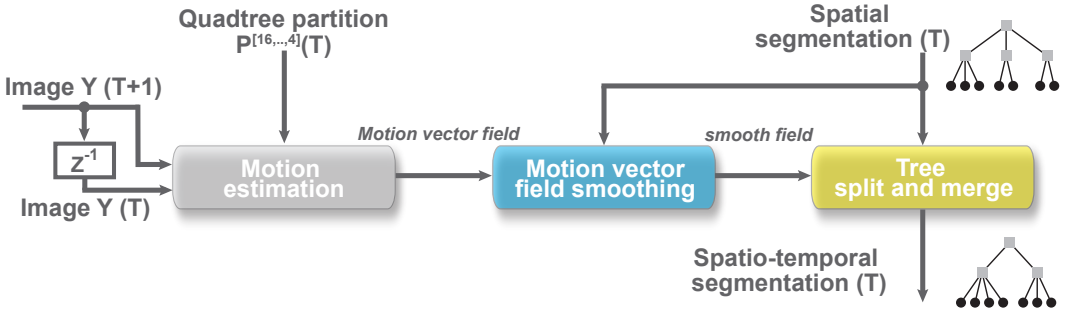


Figure 6.4: Spatio-temporal segmentation process: motion and spatial compliant hierarchy

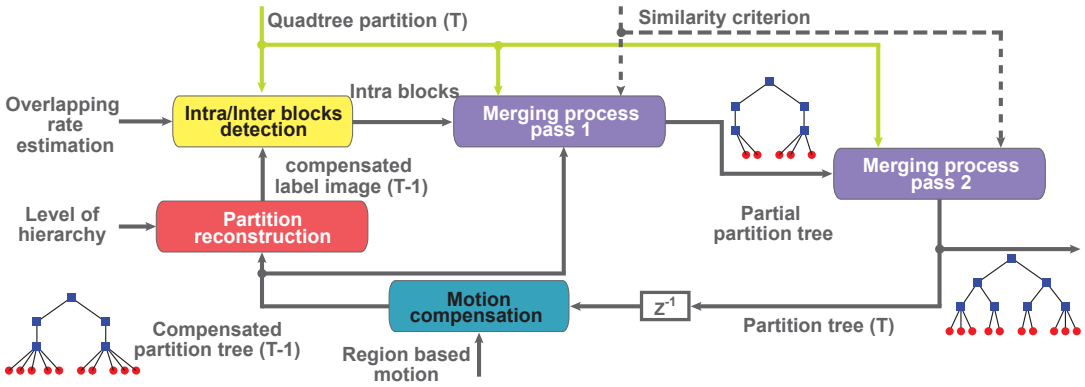


Figure 6.5: Spatio-temporal segmentation process: motion and spatial compliant hierarchy

(a) Quadtree partition

(b) Decoded image + post-processing

(c) Segmentation with 120 regions

(d) RAG

Figure 6.6: From a moving patch of texture towards the motion tube

an interpolation process. Finally, once a level of the hierarchical spatio-temporal segmentation has been selecting, motion vectors associated to the regions are transmitted.

6.1.1.3 Temporal consistency of region representation

To obtain a consistent region based video representation, different tools have been adapted to our framework. Based on a two-step merging process, labelling consistency is preserved so that to insure a temporally stable representation. As shown on figure 6.5, previous partition tree is projected onto quadtree of the current frame. Then, depending on the quality of block motion compensation, the label of the previous region representation is affected to the current block or not. The next step only considers the well-preserved blocks and applies a dedicated segmentation step. As for it, the second segmentation pass processes the remained blocks thus providing a complete multi-scale region representation.

6.1.2 Results and discussion

In this section, some visual results are described. Resulting from the Foreman CIF encoding with a bit-rate of 350kbit/s, figure 6.6b shows the decoded P-frame 50 with a PNSR equal 29/36/36 dB for respectively Y/Cr/Cb components (4:2:0). The luminance is encoded with the Flat Interleaved S+P (inter/intra mode) as shown in section 6.1.1.1 and the chromatic component representation (120 regions) can be described with the region adjacency graph (RAG) presented by the figure 6.6d.

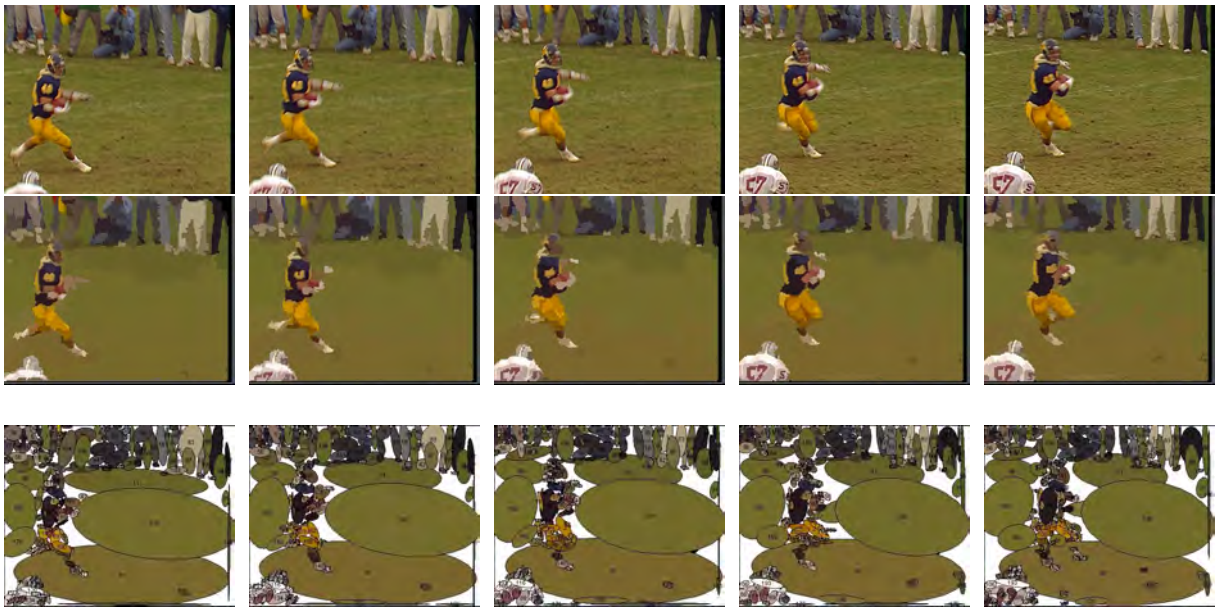


Figure 6.7: Temporal consistency of regions: illustration on *football* sequence. First row shows the original sequence, second raw the associated region representation, third row illustrates region labeling

Though compression results remains lower than the standard MPEG4-AVC, offered functionalities by this multi-scale representation with a controlled coding cost are very interesting.

The temporal consistency of the representation is illustrated by figure 6.7. The third row of images shows ellipses that correspond to regions obtained by the framework. The ellipse centers match the centroïd locations of each region, and their size is proportional to the size of considered regions. This figure illustrated the consistency as the region labels remain persistent throughout the sequence.

Classically, region-based video coder does not provide competitive solutions regarding the lone bitrate. In particular, the forward motion estimation, required in case of region based motion description, drastically lower the coding efficiency. However, we demonstrate here the ability of the framework to both provide a temporally consistent region representation allowing region matching/tracking together with a contained complexity. Considering non-explicit Partition Tree encoding, the proposed framework provides a solution able to provide multi-scale content representation and thus enabling advanced functionality such as ROI coding and video object tracking.

To match coding efficiency requirements, region-based solutions are then not relevant. To alleviate this issue, joint block-based and texture-based solutions have been designed. The following section presents an innovative motion tube representation.

6.2 Motion tubes representation for image sequences

While looking at a sequence of natural images, one can see that a texture is likely to be found in several consecutive frames. Indeed, the textural information is carried by the objects of the scene and its the background: both objects and background are most often persistent through the time.

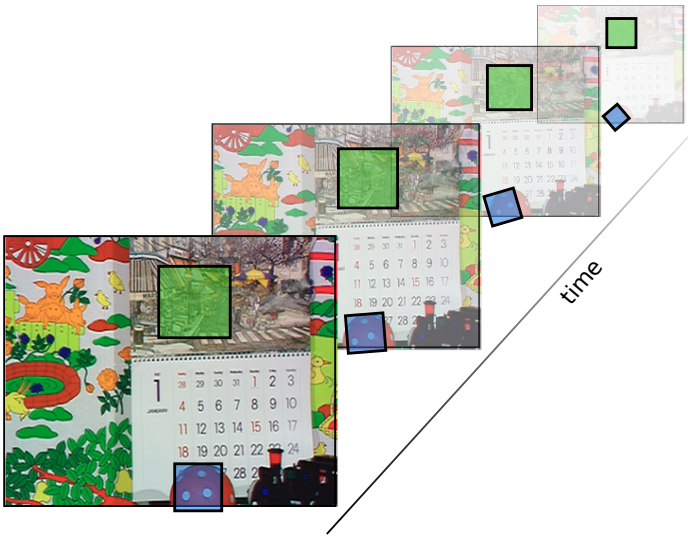


Figure 6.8: Temporal persistence of textures in image sequences

However, due to camera motion and object displacement, any texture may follow a given trajectory and may undergo a certain deformation (translation, resizing, rotation, warping) as shown on figure 6.8.

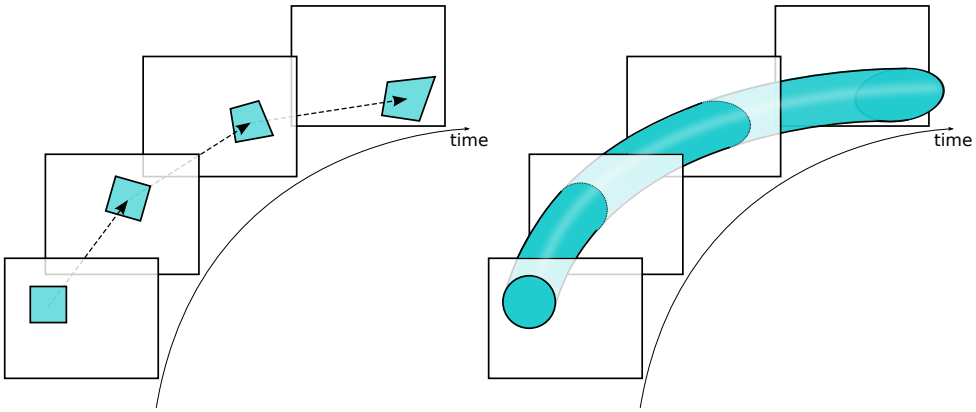
As a consequence, any sequence of images can generally be considered as a set of moving patches of texture, with respect to a given set of trajectories and deformations. From that perspective, it is reasonable to represent a sequence of images as such, assuming that it is possible to get a correct set of patches along with their trajectories and deformations parameters. Some textures, however, cannot be handled by such an approach: particle objects, liquids and transparencies, to cite a few of them, present such major changes of texture that it might be impossible to track their evolution over time without any specific approach. The current work focuses on sequences which do not present such delicate textures, which will be processed as for them in section 6.3, and aims at finding a set of patches of texture to represent such sequences.

It appears that an ideal compression system might be obtained by locally decorrelating the spatial contents, and globally processing the information along the temporal axis. In practice, this can be obtained by continuously tracking local areas of an image sequence across time. Putting aside the traditional frame-*macroblock* paradigm, it is proposed to see image sequences as a collection of spatio-temporal units which deform and move across time and space. These structures are called motion tubes, and are described in this section.

This work is the result of a collaboration with Stéphane Pateux and Nathalie Cammas (Orange Labs), who co-supervised Matthieu Urvoy during his PhD work [196][195].

6.2.1 Modeling a motion tube

As depicted on figure 6.9, the successive shapes of a moving texture patch reminds a tube whose section is deforming. From now on, moving patches of textures will be referred to as *motion tubes*. Such motion tubes will be able to track moving texture across time and space, thus exploiting their temporal persistence.



(a) A moving patch of texture (b) A crude representation of a motion tube

Figure 6.9: From a moving patch of texture towards the motion tube

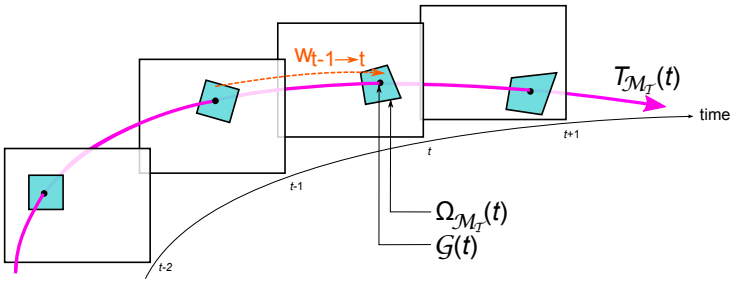


Figure 6.10: Trajectory and deformation of a motion tube

Tubes definition. Let $\mathcal{M}_\mathcal{T}$ be a motion tube attempting to describe a patch of texture along with its evolution across time and space, through a sequence of images. $\mathcal{M}_\mathcal{T}$ will be characterized by 3 types of information: its texture $\boldsymbol{\mathcal{T}}$, its lifespan $\boldsymbol{\mathcal{L}}$, and its deformation parameters $\boldsymbol{\mathcal{W}}$ leading to

$$\mathcal{M}_\mathcal{T} = \{\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{W}}\} \ . \tag{6.1}$$

Deformation model. In order to cope with texture displacements and deformations, an appropriate deformation model has to be set up. The deformation of a texture may be described as the result of a transform whose transfer function is a *warping* operator w.

Trajectory. The notion of trajectory is crucial to the notion of motion tubes. Indeed, the temporal persistence of textures will be captured only if motion tubes naturally exhibit a deformation along a trajectory $T(t)$. A luminance sample $\boldsymbol{\mathcal{P}}$ of the patch of texture, due to the displacement and the deformation of the texture, will follow a trajectory $T_{\boldsymbol{\mathcal{P}}}(t)$. The trajectory $T_{\mathcal{M}_\mathcal{T}}(t)$ of a motion tube will be given by the trajectory of the center of gravity $\mathcal{G}_{\mathcal{M}_\mathcal{T}}(t)$ of the patch at each time instant t. Figure 6.10 illustrates this principle.

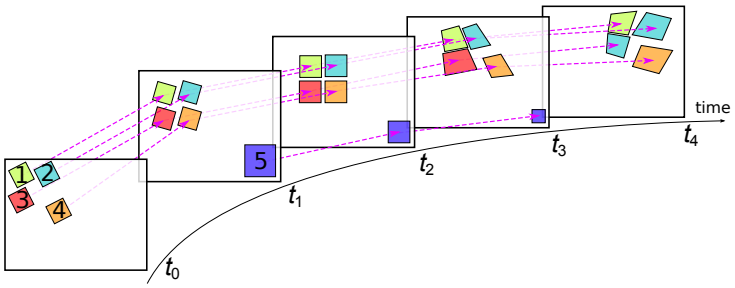


Figure 6.11: An image sequence partially reconstructed from a few motion tubes

6.2.2 Video representation based on motion tubes

Figure 6.11 illustrates a motion tube based representation. Five tubes have been initialized at $t_{t_{\text{start}}} = t_2$, and their textures have been tracked from t_0 to t_4. Assuming the motion field is uniform on areas reconstructed by motion tubes 1, 2 and 3, we can see that they are kept together through the whole GOP. Tube 4, on the other hand, behaves differently to map a discontinuity of the motion field. Tubes 1, 2, 3 and 4 share the same start ($t_{t_{\text{start}}} = t_0$) and end ($t_{t_{\text{end}}} = t_4$) instants, and their lifespan $\boldsymbol{\mathcal{L}}$ is $[t_0, t_4]$. Finally, tube 5 does not appear at t_0 nor t_4 because the texture it carries is not present at those instants: its lifespan is then $[t_1, t_3]$.

This being said, the main problem of this representation comes down to find a right set of motion tubes. Let I_t be an image from the sequence $\mathcal{S}_{\mathcal{I}}$ at time instant t. While the existence of a matching set of motion tubes S can be easily assumed, there is no uniqueness for such a representation.

6.2.2.1 Motion tubes families

A motion tube is driven by a large set of parameters (lifespan, shape, trajectory, texture, ...). In practice, its optimization may prove to be a difficult task. In order to simplify this problem, it is proposed to create groups of motion tubes which, for now, will share the same temporal parameters. These groups are called *families* of motion tubes.

A sequence of images represents a scene within which, due to camera motion or objects displacements, background may change and objects may appear or disappear. Likewise, the availability of the textures will also vary across time. A family of motion tubes sources the textural information from a common reference instant. As a consequence, it will not be able to register the entire textural information: a single family cannot entirely represent a complex sequence.

Therefore, several families of motion tubes will be required to provide an appropriate representation. These families may overlap temporally and/or spatially. A convenient way to instantiate several motion tube families is to split the sequence into GOP and create a family for each of them. Each family is initialized at the GOP start instant. This particular solution is illustrated in figure 6.12 using GOP of 8 time instants. The first GOP is reconstructed with the red family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_{\mathcal{T}}}(t_0)$ whose tubes end at t_8, the second GOP by the blue family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_{\mathcal{T}}}(t_8)$ whose tubes end at t_{16}, and the third GOP is reconstructed by the green family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_{\mathcal{T}}}(t_{16})$ which ends at t_{24}. Typically, camera motion and objects displacement may force us to create several families within the same GOP.

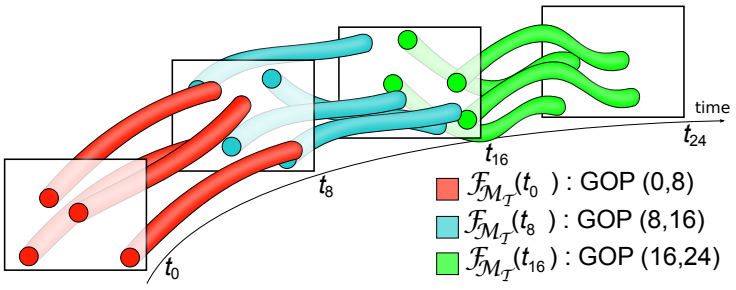


Figure 6.12: GOP paradigm in the context of motion tubes

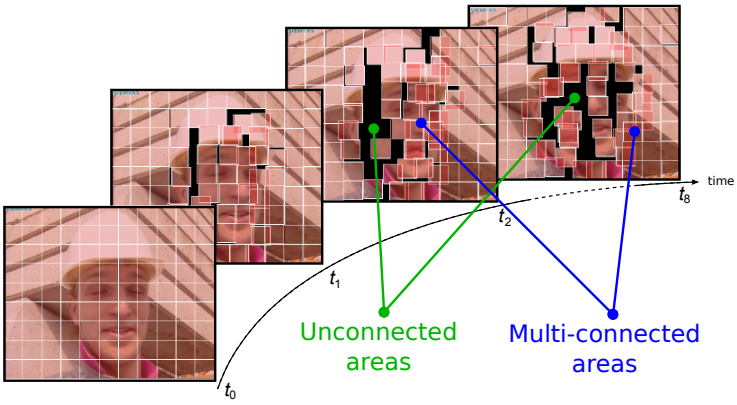


Figure 6.13: Preliminary example of motion tube based reconstruction - Connected and unconnected tubes

6.2.2.2 Motion tubes video representation properties

Motion tubes, by their nature, can benefit from the temporal persistence of moving textures: their tracking is then simplified. In particular, the motion estimation of a motion tube, at a given instant, can be guided by its trajectory at previous or next instants. This tends to reduce the discrepancies of the motion field, and the relative motion coding cost. It also maximizes the tube's lifespan, due to an enhanced tracking, thus minimizing the amount of textural information to be sent. Furthermore, as motion tubes can start and end at any time instant, they fit appropriately the instants of apparition and disappearance of the tracked textures. They can be either dependent or independent from each other (neighbouring motion tubes undergo the same changes), *connected* or *disconnected* (neighbouring motion tubes may be joint or disjoint). Keeping motion tubes connected will constraint the motion field to be continuous, while disconnections will be able to represent the ruptures of the motion field.

Sole motion tubes might not be able to entirely represent a sequence of images. In that, they may be seen as a synthetic representation. In particular, complex scenes on which the tracking of textures is difficult or even impossible won't be entirely reconstructed by any reasonable set of motion tubes. In figure 6.13, unconnected areas correspond to areas which are not reconstructed by any motion tube. Dedicated mechanism will then need to be proposed to handle these areas (*e.g.* with inpainting).

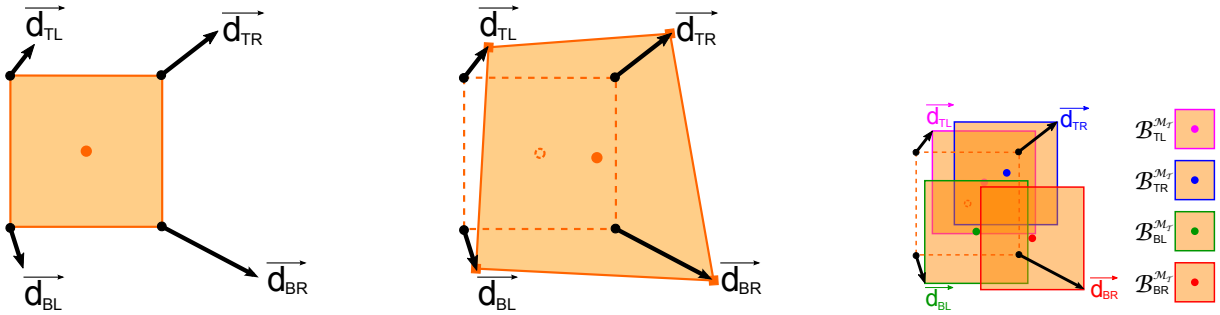
6.2.3 Motion model of a tube

Among pattern-based models, block-based and mesh-based models are very popular in video compression applications. Historically, however, blocks are generally preferred to meshes as they are used in standardized video compression schemes. Yet, numerous techniques employing blocks and/or meshes have been provided. An advanced study of these techniques has been realized in [195] considering our context. The ability of several models have been studied, in particular Block Matching Compensation algorithms (BMC) [190][87], Overlapped Block Motion Compensation (OBMC) [136][140][33], Control Grid Interpolation (CGI) [186] and a few hybridized variants such as Switched Control Grid Interpolation (SCGI) [76], Switched OBMC (SOBMC) [77].

If hybrid CGI-based approaches generally provide the best abilities regarding motion representation, they do not fulfill our complexity requirements. As a consequence, SOBMC provides the best trade-off between modeling abilities and complexity and is considered as the basement of our proposed motion model. However, the SOBMC has been only used to model the deformation between a couple of images, and does not provide (unlike active meshes, for instance), an intrinsic ability to describe the deformation of a whole GOP in a continuous manner. As a consequence, the provided motion model will extend the abilities of the SOBMC to account for the spatio-temporal nature of the motion tubes.

6.2.3.1 In between blocks and meshes: a modified Switched OBMC motion model

In order to represent a wide enough variety of deformations, the motion of each motion tube $\mathcal{M}_\mathcal{T}$ is described through four motion vectors. To each corner of the quadrilateral patch of texture being tracked is associated a motion vector which describes its displacement in regards to its initial position in the reference image. At time instant t_n, $\overrightarrow{d_{\text{TL}}}(t_n)$, $\overrightarrow{d_{\text{TR}}}(t_n)$, $\overrightarrow{d_{\text{BL}}}(t_n)$ and $\overrightarrow{d_{\text{BR}}}(t_n)$ respectively stand for the displacement of the top-left, top-right, bottom-left and bottom-right corners of $\mathbf{\Omega}_{\mathcal{M}_\mathcal{T}}(t_n)$.



(a) Initial block of texture $\mathbf{\Omega}_{\mathcal{M}_\mathcal{T}}(t_{\text{ref}})$

(b) Idealistic forward CGI warping

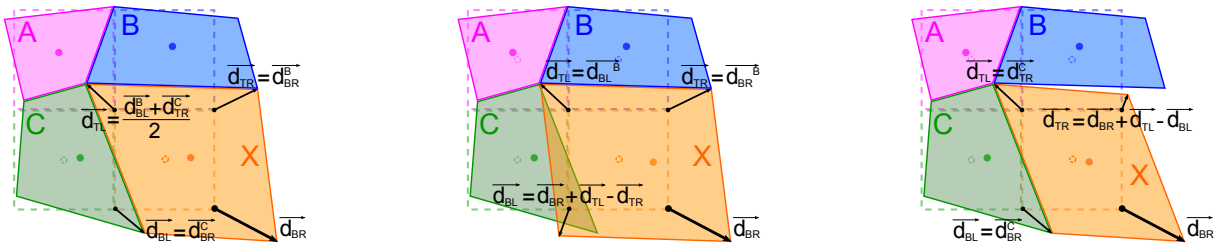
(c) OBMC-like warping approximation

Figure 6.14: Forward motion compensation of a motion tube: in between OBMC and CGI

The proposed deformation model is illustrated in figure 6.14. Figure 6.14a shows the initial block of texture sourced from I_0 along with its motion vectors. As each motion vector describes the local displacement of a single corner, the proposed model can be compared to the CGI, wherein all four corners are behaving as control points: the original square block can then be seen as a mesh whose deformations motion compensate the corresponding patch of texture (figure 6.14b).

6.2.3.2 OTMC: connected/disconnected motion tubes

Connected motion tubes. In order to provide a continuous representation of the motion field, it is crucial for the motion tubes to be able to remain connected to each other, and for their motion model, to handle corresponding deformations (warpings). Let **X** be the current motion tube. Let **A**, **B** and **C** be respectively the top-left, top, and left causal neighbours of **X**. It is assumed that their deformation has been previously estimated.



(a) Full-connected OTMC (b) Top-connected OTMC (c) Left-connected OTMC

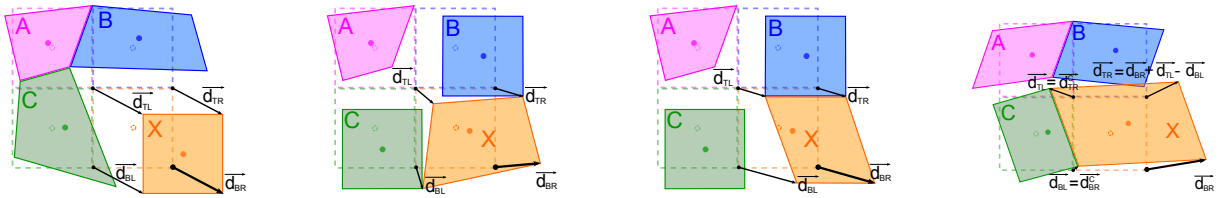
Figure 6.15: Idealistic representation of the deformation for the three connected modes

The proposed OTMC model keeps **X**'s corners connected to those of its causal neighbours. In practice, vertical and horizontal connection directions are only considered. Figure 6.15 illustrates the OTMC motion model. In order to keep the schematics as simple as possible, the deformations are represented as idealistic CGI warpings. Three connection modes are provided: full-connected OTMC mode, top-connected OTMC mode and left-connected OTMC mode.

TMC: disconnected motion tubes in translation. Despite its crudeness, the translational BMC motion model proved to be quite efficient in classical approaches to motion compensation, in terms of compression. In particular, its ability to compactly represent the discontinuities of the motion field have been largely appreciated. Similarly, it is crucial for the motion tubes to be able to simply translate, regardless to the deformation of neighbouring patches of textures. In such case, a single motion vector (in practice, $\overrightarrow{d_{BR}}$) is used to describe the translation undergone by the motion tube. Such a motion model is baptized TMC and requires then a single translational projection to be performed.

Hybridizing TMC and OTMC motion models into a Switched OTMC model Similarly to the SOBMC and other hybrid block-mesh approaches, the final motion model hybridizes the different motion modes previously introduced. Four connection modes are now available: full connection, left connection, right connection and disconnection. The different motion modes can be hybrized in various ways to accurately represent the local variations of the motion field. Figure 6.16 illustrates their hybridization with several connection patterns which may appear.

With TMC and OTMC motion modes, neighbouring motion tubes can be either fully connected, or completely disconnected. At some point, however, it may be interesting to partially connect a motion tube to only one of its neighbours. As a solution, two intermediate motion modes lying in between the disconnected TMC and the connected OTMC motion modes are proposed:



(a) **A**, **B** and **C**: OTMC; **X**: TMC (b) **A** and **X**: OTMC; **B** and **C**: TMC (c) **A**: OTMC; **B** and **C**: TMC; **X**: top OTMC (d) **A** and **C**: OTMC; **B** and **X**: left OTMC

Figure 6.16: Various hybridizations of the different motion models

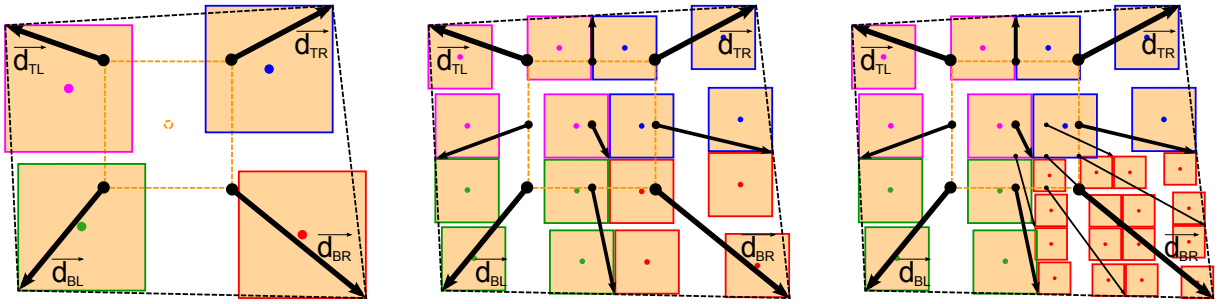
- the Top OTMC motion mode keeps the current motion tube **X** connected to its top neighbour **B** (figure 6.16c);
- the Left OTMC motion mode keeps the current motion tube **X** connected to its left neighbour **C** (figure 6.16d).

6.2.3.3 Regularizing the motion discrepancies: Locally-Adaptive OTMC

While OTMC provides the ability to connect neighbouring motion tubes, it represents the deformations in a very crude way: indeed, any geometric shape is reduced to a set of four overlapping blocks $\mathcal{B}^X_{TL}$, $\mathcal{B}^X_{TR}$, $\mathcal{B}^X_{BL}$ and $\mathcal{B}^X_{BR}$. This crude representation may not be appropriate to any kind of deformations, in particular in case of large deformations. This results into motion discrepancies which could be avoided with a finer motion model.

As a solution, it is proposed to recursively split these four blocks into sub-blocks. Their displacements are interpolated from the four original motion vectors. This process is iterated until the representation of the deformation is accurate enough. Such a motion compensation process is called LAOTMC, and is illustrated in figure 6.17.

The decision whether to split or not to split a (sub-)block relies on its four displacement difference between consecutive corners. If all the differences are under a coherence threshold δ_{Th}, the partitioning operation is not required, and the (sub-)block can be directly motion compensated.



(a) Initial OBMC blocks (b) First level partitioning (c) Second level partitioning

Figure 6.17: LAOTMC motion mode: automatic recursive partitioning of a motion tube

As can be seen from figure 6.17, the accuracy of the deformation representation is locally adapted to the nature of the motion field. In particular, it can be seen from figure 6.17c that the shape's

approximation is much more closer to the idealistic shape (represented by the dashed quadrilateral) than its initial crude approximation was. Using the LAOTMC, a larger set of deformations can be then handled.

6.2.3.4 Motion modes: compared performances

Disconnected TMC versus connected OTMC. In [195], it has been shown that TMC motion mode provides a significantly better PSNR on reconstructed areas (SPSNR metric) than the OTMC does. On the other hand, the OTMC is able to reconstruct a larger proportion of the images. This is easily explained by the fact that the OTMC motion mode is not able to represent the discontinuities of the motion field, and cannot catch up with areas corresponding to moving objects boundaries. However, forcing the motion tubes to be connected to each other limits the amount of unpredicted areas. As for it, figure 6.18 shows the motion compensation prediction of the third frame I_2 of sequence *Foreman* whether TMC or OTMC motion modes are used. Compared to the TMC, the OTMC does increase the amount of reconstructed areas, while lowering the overall quality of the compensation. In particular, distorted edges can be observed.

(a) Original frame

(b) TMC - SPSNR= 32.55 dB, Rec. rate= 91.10%

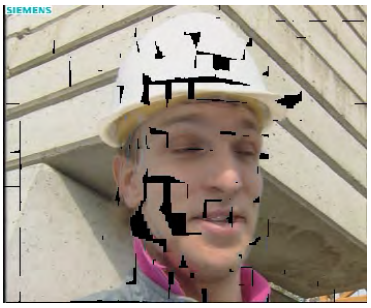
(c) OTMC - SPSNR= 29.80 dB, Rec. rate= 99.03%

Figure 6.18: TMC versus OTMC motion modes: influence on the synthesized images

How hybridizing the different motion model does improve the motion compensation. While the efficiency of the OTMC in regards to the TMC is arguable, their hybridization undoubtedly ends up in an improved motion compensation. They all provide PSNR higher than the reference score of the TMC. Hybrid scenarios, however, do not provide a reconstruction rate as high as the sole use of the OTMC does. In the end, hybridizing all four motion modes respectively increases both PSNR and reconstruction rates by 1.82 dB and 2% in average. Figure 6.19 shows the motion compensation prediction of the third frame I_2 of sequence *Foreman* obtained with three hybrid scenarios. The full OTMC motion mode is advantageously used in areas which contain few edges. On the contrary, edgy areas are most often handled by partially disconnected (top and left OTMC) or disconnected (TMC) modes. If all four motion modes are enabled, the selected motion models roughly divide up into: 40% of TMC, 30% of full OTMC, 15% of left OTMC and 15% of top OTMC as well.

(a) Reconstruction of I_2 using full OTMC, left OTMC and top OTMC modes
SPSNR= 32.93 dB, Rec. rate= 96.91%

(b) Reconstruction of I_2 using TMC and full OTMC modes
SPSNR= 33.47 dB, Rec. rate= 95.78%



(c) Reconstruction of I_2 using TMC, full OTMC, left OTMC and top OTMC modes
SPSNR= 33.96 dB, Rec. rate= 95.44%

Figure 6.19: Hybridization of the four motion modes: influence on the synthesized images

6.2.4 Time-evolving representation of the textures

The evolution of a patch of texture across time is not only defined by its displacement and deformation, but also by its intrinsic textural changes. Towards an accurate time-evolving representation of the textures, it is proposed to source the textural information from several reference images. In such case, the final texture results from a weighted average of several temporally-aligned patches of textures. Motion tubes benefiting from several texture reference instants are called *bi-predictive tubes*, or shortly, *B-tubes*. Let $\{I_0, \ldots, I_8\}$ be a GOP of $G_S = 9$ consecutive images. From now on, it is assumed that two reference images $I_{\textbf{ref0}}$ and $I_{\textbf{ref1}}$ are available. Typically, these can be located at the extremities of the GOP, such that $I_{\textbf{ref0}} = I_0$ and $I_{\textbf{ref1}} = I_8$.

B-tubes definition. *B-tubes* source their textural information from more than one reference instants. In this way, existing video compression schemes based on Analysis-Synthesis framework similarly use several reference instants to synthesize the images [188, 198, 29, 103]. All these video compression schemes, however, suffer from resolution losses, as they use a single reference instant in the end: textures from the many different reference instants are first projected on a *main* reference instant; the different contributions are then merged and projected into the current instant to obtain the final synthesized image. As textures undergo two successive motion compensations, projected textures often suffer from resolution losses. Motion tubes, however, do not suffer from this problem, as there is no *main* reference instant. Textures are in our case motion compensated a single time only, which reduces the risks of resolution losses, as motion tubes warping operators could be easily composed and inverted, such that any image from any instant can be directly motion compensated at any other time instant.

B-families. At this point, motion tubes mainly suffer from their inability to provide a complete reconstruction of the images. Small and large holes were previously distinguished. Large reconstruction holes are especially problematic, as they generally correspond to textural areas which are unavailable at the reference instant. They can be the consequence of occlusions, camera motions, etc. As a solution, it is proposed to include additional motion tubes to describe these areas. Obviously, these

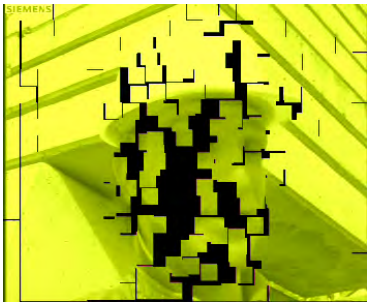
motion tubes need to be instantiated from a frame which actually contains the textures missing to the spatio-temporal representation. The concept of *B-families* of motion tubes has been then introduced, as an inter-tube bi-prediction mechanism.

Figure 6.20 shows how *B-families* of motion tubes drastically improve the representation. Top figures show the output images, while bottom figures indicate which families of motion tubes have contributed to the reconstruction: yellow for $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0)$, blue for $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_8)$, green for a weighted average of $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0)$ and $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_8)$, and finally magenta for $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_4)$. As could be expected, family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0)$ mainly contributes to the reconstruction of the right half of the face of Foreman, while family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_8)$ mainly contributes to the reconstruction of its left half. Both of these families have been able to catch the deformation of the background, which is mostly synthesized from a weighted average of both contributions. Finally, the third family $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_4)$ nearly completes the reconstruction, by notably providing a large part of the mouth which was missing.

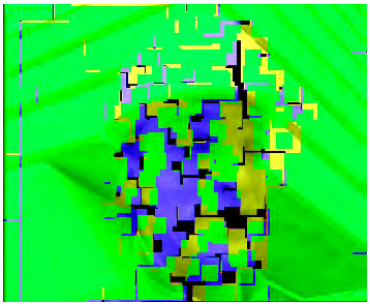
(a) $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0)$'s reconstruction SPSNR= 35.79 dB, Rec. rate= 86.81%

(b) ${}^0\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0,t_8)$'s reconstruction SPSNR= 36.52 dB, Rec. rate= 96,26%

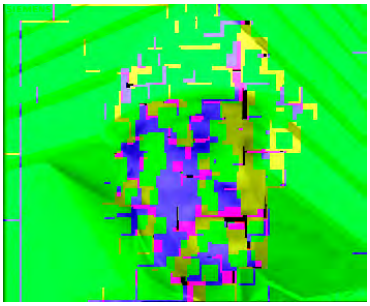
(c) ${}^1\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0,t_4,t_8)$'s reconstruction SPSNR= 36.24 dB, Rec. rate= 99.29%



(d) Areas reconstructed by $\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0)$



(e) Areas reconstructed by ${}^0\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0,t_8)$



(f) Areas reconstructed by ${}^1\boldsymbol{\mathcal{F}}_{\mathcal{M}_\mathcal{T}}(t_0,t_4,t_8)$

Figure 6.20: *B-families* of motion tubes: influence on the synthesis of the fourth image of sequence *Foreman*

6.2.5 Representation and compression ability of motion tubes

Video coding: integration into AVC standard. All motion tubes were systematically tracked during the whole duration of the GOP, whether the corresponding patches of textures could be tracked or not. Occlusions, complex deformations, or any other phenomena preventing an initial

patch of texture from being tracked are especially problematic for the motion tubes. It is critical for the representation to be able to detect their occurrences and, in that event, deactivate or terminate the affected tubes.

In order to further validate the representation in regards to state-of-the-art coding tools, it was first proposed to compare the reconstruction provided by the motion tubes with the reconstruction provided by AVC compression standard. To this end, motion tubes were embedded into the JSVM. The decisions taken by the modified coder provided a large amount of invaluable informations regarding the motion tubes and their abilities. A complete study can be found in [195]. Several key features were then highlighted:

- motion tubes were eventually used, in very significant proportions, by the modified coder. This further validates the concept of motion tubes: they are able to provide a proper representation of a large part of the sequences;
- as could be expected, the more stable and persistent the images contents are across time, the more suited are motion tubes. Complex areas from the sequences, however, were generally better matched by classical coding tools.
- in terms of pure compression, however, the proposed hybrid compression scheme does not provide any gains; quite the contrary, actually. Still, the coding cost of the motion tubes can be easily reduced by removing the motion information of unused tubes.

The integration of the motion tubes into AVC showed that a significant amount of the macroblocks could be favorably coded by the motion tubes. However, the coding cost of the motion tubes was not taken into account in the mode selection process, as the instantaneous coding cost of a motion tube has little if no meaning. In the end, the gain in pure AVC bitrate dit not compensate for the coding of the motion tubes. On the other hand, all motion tubes were transmitted, even though they were only partially or not selected at all by the modified coder. For this reason, it is essential for unused motion tubes to be discarded from the representation, thus reducing their coding cost. Compression performances may then turn in favor of the motion tubes.

Representation features. During his PhD work, Matthieu Urvoy has proved that motion tubes are actually viable candidates to the representation of image sequences. Reconstructed images showed that, for a large proportion of the image sequences, motion tubes were able to properly synthesize the textures from a minimal amount of textural information. Their integration to H.264/AVC further proved how much interesting motion tubes may be in regards to more classic representations. As expected, they also intrinsically exhibit the temporal persistence. In terms of representation, this turns out to be an invaluable information regarding the spatio-temporal content of the image sequences, and may greatly facilitate its analysis. Even though initial patches of textures cannot be interpreted as semantic regions, their evolutions will submit to those of the objects they belong to. A video object may and its temporal evolution may then be interpreted as an arbitrary set of motion tubes.

Motion tubes were then designed in such a way that the proposed representation is not affected by problems which are inherent to existing schemes. However, such a disruptive approach comes to a cost: the tube-based representation is not mature enough yet to be advocated for practical use for video coding purposes (*e.g.* as an answer to calls from normalization committees). Also, its ability to compact image sequences needs to be significantly increased for it to be seriously considered as

an alternative. Nevertheless, the pseudo-semantic representation ability of the motion tube can be seen as an innovative analysis tool.

6.3 Adaptive image synthesis for video coding

In state-of-the-art compression schemes, pixel-wise redundancy is reduced by using predictions and transformed domain operations. However, classical spatio-temporal approaches, exploiting redundancy based on the mean squared error (MSE) criterion, are not able to take visual redundancy into account. Detailed textures may seem nearly stationary for the Human Visual System, but totally irregular according to MSE criterion. At the same time, texture synthesis algorithms [201, 6, 96, 95] have shown promising results. The purpose of new coding schemes is to detect regions where exact positions of texture patterns are irrelevant for the human eye. The whole regions are not encoded since a few patterns are sufficient enough to synthesize satisfactory regions.

The following framework has been proposed by Fabien Racapé during his PhD work [151], thanks to a industrial collaboration with Edouard François, Dominique Thoreau and Jérôme Viéron from Technicolor [152][153].

6.3.1 Motivation

One of the first synthesis-based compression scheme was presented in [48]. This approach includes an analysis in order to detect replaceable textures which are finally synthesized by a dedicated texture synthesis algorithm at decoder side. An interesting framework has been designed in [220] where some 8x8 blocks are removed at the encoder side and synthesized by the decoder. The segmentation first classifies blocks into structures, corresponding to boundaries of objects, and textures. The first category is classically encoded and the latter is removed and synthesized at decoder. The segmentation is based on simple edge detector thresholding. To avoid temporal inconsistencies, motion estimation is considered in selecting patches which will be used for texture synthesis. The latter is performed using the algorithm presented in [96], which appeared hard to exploit in our tests, since 8x8 blocks offer poor overlap for the Graphcut technique.

The work presented in [134] proposes a closed-loop analysis-synthesis approach. As in [220], groups of pictures (GOP) are considered for a spatio-temporal scheme. Each potential region from a GOP is both analyzed and synthesized, using side information from texture analyzer. A first synthesizer is designed for *Rigid textures* with global motion, which has a great similarity to global motion compensation (GMC). Another synthesizer, inspired by the patch-based approach developed in [96], processes *Non-rigid textures* with local and global deformations. The scheme has recently been upgraded [133] with photometric corrections, using Poisson editing and covariant cloning. However, Y. Liu [111] pointed out that synthesis algorithms work more efficiently with different kinds of textures and addressed the specific case of near regular textures.

Another framework for image compression is proposed in [109] where both textural and structural blocks can be removed at encoder side. The textures are synthesized at decoder using [96], while a binary edge information helps an inpainting method to retrieve structural ones. To keep color coherency, a draught board of blocks is preserved, limiting the bit-rate saving. The problem of these gradated regions is tackled in [213] with parameter assistant inpainting. However, several pixel-based synthesis algorithms from the literature [201, 6, 95] also provide promising visual performance for a

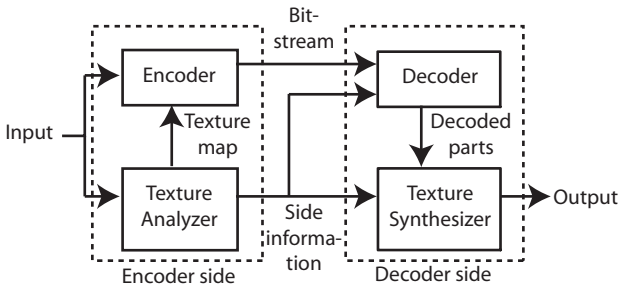


Figure 6.21: Framework overview.

large range of textures, until they get the right parameters depending on the patterns to synthesize.

An innovative framework using both complementary pixel-based and patch-based approaches has been then realized. The scheme includes a texture characterization step, based on DCT-domain descriptors, that outputs the approximated feature size of texture patterns to be synthesized. The two following texture synthesis algorithms have inspired this work: the basic pixel-based approach developed in [201] and the algorithm presented in [96] for the patch-based synthesis.

The proposed framework is depicted in figure 6.21. First, input images are analyzed to detect textured regions that can be synthesized by the algorithm used at decoder side. Thus this part of the framework is divided into two main steps: a segmentation step and a characterization step. A resulting texture map is sent to the encoder and side information, describing texture patterns. Then, pointed out regions are partially removed, whereas structural regions are classically encoded. Removed regions are synthesized from neighboring small surfaces of texture, also classically encoded. In the following, the patches denote these samples of texture which serve as inputs for synthesis. At the decoder side, the bit-stream and potential side information to locate encoded regions are both used to build and locate structural parts, whereas the labeled removed regions are synthesized using a new adaptive pixel-based algorithm. Like encoder side analysis, the synthesis contains a characterization step, which outputs required parameters for texture algorithm.

6.3.2 Texture analysis

This section focuses on the analysis at the encoder side which aims at characterizing texture patterns and designing the removed regions and sample patches for synthesis. For validation purposes, the segmentation has been performed by the region extraction scheme presented in section 2.2.3.

6.3.2.1 Texture characterization

This part of the scheme is used at both encoder and decoder sides. It outputs the parameters required by the texture synthesizer in order to give its best results. Thus, using characterization enables the encoder to decide whether the texture can be synthesized or not. Indeed, if the characterized texture requires impossible parameters, for example too large patterns, it is finally classically encoded. The main parameter for our pixel based algorithm corresponds to a size of neighborhood window. Thus the characterization step computes descriptors from different sizes. The descriptors are derived from those described in [178] which are computed from the Fourier transform. Since the characterization requires little sizes of blocks, i.e. inferior to 32 pixels wide, descriptors from DCT domain have been

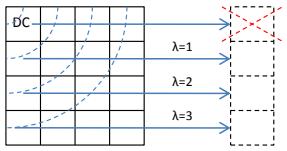


Figure 6.22: Descriptor computed on a 4x4 DCT block.

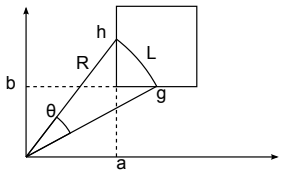


Figure 6.23: Descriptor Integral on a discrete array.

chosen. Indeed, the Fourier transform outputs descriptors with a lower resolution in frequency, which is a problem to describe a signal from a small window. Like in [178], descriptors are computed from concentric circles represented in figure 6.22.

The DC coefficient is not taken into account in order to be invariant to changes of average luminance, so the descriptors vector has size of block minus one coefficient. A large set of blocks inside the texture region are randomly chosen to compute descriptors at different sizes centered at the same position.

Since the scheme uses discrete transform, figure 6.23 depicts the integral computation on a particular coefficient at position (i, j). The DCT value is weighted with the length L of the arc of a circle, crossing position (i, j) from a to b, given by $L = R * \theta$ where

$$\theta = arcsin\left(\sqrt{1 - \frac{b^2}{R^2}}\right) - arcsin\left(\frac{a}{R}\right). \tag{6.2}$$

The average descriptors are then computed for each size and analyzed. According to [201] and our experiments, the size of the compared neighborhood has to be greater than the elementary pattern of the texture, to produce a visually good result. The first coefficient of each block corresponds to a single variation of luminance over the block used for DCT computation. Experiments on a large set of texture patches show that if the first coefficient is greater than any others in the descriptors vector, the block size is smaller than an elementary pattern.

Thus, computing descriptors at random locations will serve to approximate the feature size of texture patterns. Typically, descriptors of sizes from 4x4 up to 16x16 pixels are computed for each location.

The next section describes the synthesis of the segmented regions, using parameters from texture characterization.

6.3.3 Texture patch design

Texture synthesizers from the literature process with a rectangular input patch. Its relevance is essential to ensure an efficient synthesis in this particular context. One notices that cropping a patch

(a) square patch

(b) chosen patch

Figure 6.24: Texture patch design

(a) No anchor block

(b) Synthesis of an empty region

(c) chosen patch

(d) Anchor blocks location

Figure 6.25: Impact of anchor blocks to synthesized areas

close to the texture region to be synthesized can lead to inconsistencies between candidate pixels and previously reconstructed borders. Figure 6.24a shows a case in which there is a variation of luminance, leading to a failing synthesis. We propose to define the patch as the surrounding MBs of the region to be synthesized. This source patch for synthesis is depicted in figure 6.24b.

Large textured regions require anchor preserved blocks in order to prevent artefacts. After experiments, it has been decided to encode some anchor macroblocks (MB) at strategic locations depicted in figure 6.25 since the region size exceeds a fix number of MBs in width or height. Coupled with a confidence-based synthesis order described in next section, they prevent visible seams at synthesis junctions.

6.3.4 Pixel-based texture synthesis

The pixel-based synthesizer derives from the basic scheme presented in [201]. This algorithm relies on the matching between current pixels' neighborhood and those in a texture patch, using the Sum of Squares Error (SSE) as criterion. In order to improve the research process, the exact neighborhood matching described in [166] is used. The following describes the adaptation to the context of removed regions.

6.3.4.1 Adaptive neighborhood size using texture characterization

Texture characterization enables the synthesizer to get an approximate neighborhood. In order to refine the neighborhood sizes for matching, a set of neighborhood sizes are tested, for example 7x7 and 9x9 if the best descriptor is 8x8 large. The chosen distance minimizes the norm distance

$$D_{m,n}(p,c) = \frac{\sum_{k=0}^{N_{m,n}} (p_k - c_k)^2}{N_{m,n}} \tag{6.3}$$

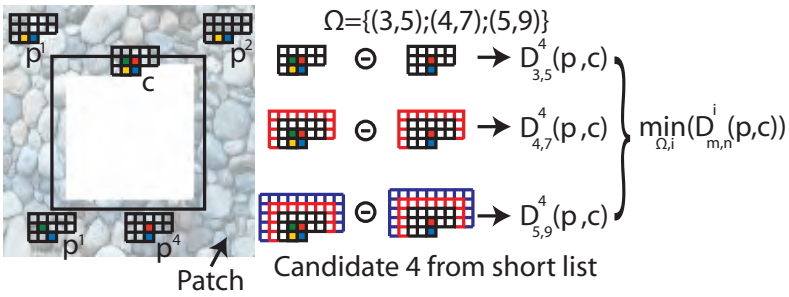


Figure 6.26: Finding best matching neighborhood.

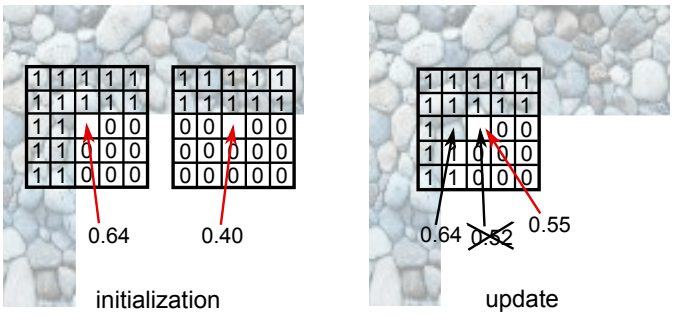


Figure 6.27: Confidence map order. Confidences w_a, w_b and w_c are computed for center pixels using neighboring confidences. At initialization, $w_a = 0.64$ for instance. After synthesizing pixel a, w_b is updated by assigning w_a^1.

where $N_{m,n}$ is the number of pixels contained in the neighborhood of current pixel, m and n respectively representing its height and width, c_k and p_k denote the kth pixel's luminance in the current neighborhood and the considered one in the patch respectively. Considering various neighborhood sizes enables to better catch texture patterns size and shape. This process is illustrated in figure 6.26 where three neighborhood sizes $\{(3;5)(4;7)(5;9)\}$ are competing. Experiments show that a synthesis with a lot of changes in neighborhood sizes does not give visually good results. In order to avoid this kind of issue, we propose to favor the neighborhood sizes that have been chosen by a majority of previously synthesized pixels, by means of a global weight.

6.3.4.2 Confidence-based synthesis order

The goal is here to exploit available data which are, at this point, the only confident data. The latter corresponds to surrounding pixels, which are whether previously decoded or synthesized. Thus, raster scan order is clearly not adapted. As in [109], a scan order depending on a synthesis-coherent confidence map is adopted. The map building law is depicted in figure 6.27, where initial computing from previously decoded pixels is represented on the left side and the map update during synthesis on the right side.

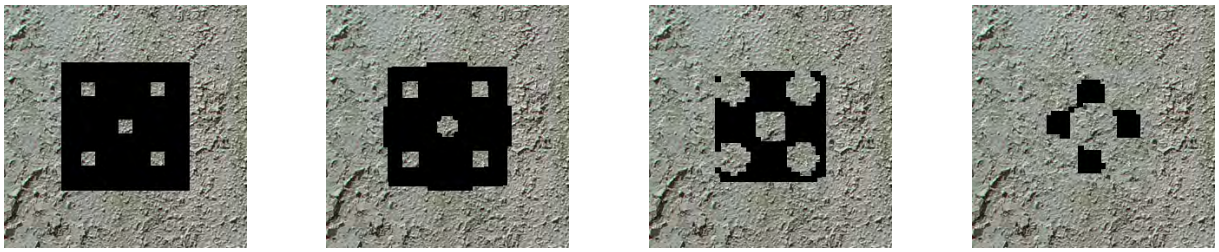
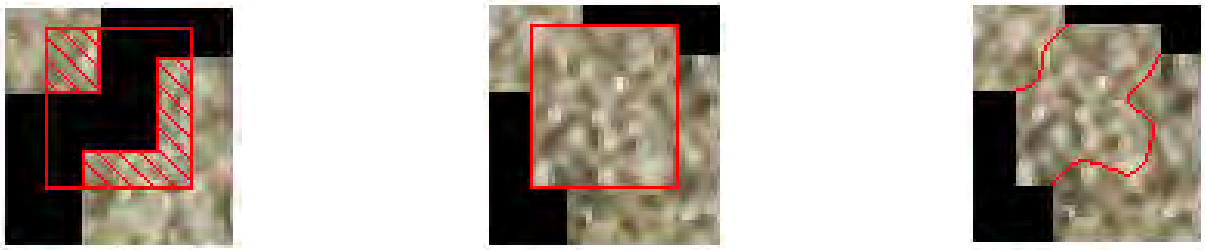


Figure 6.28: Patch-based synthesis order



(a) a (b) b (c) c

Figure 6.29: a: The red hatched region is used to search the best matching patch. b: the chosen patch is pasted on the new texture. c: the best cuts are found between overlapping regions

6.3.5 Patch based texture synthesis

The patch based method used can be decomposed into two main steps. The synthesis order, depicted in figure 6.28, is based on the confidence map. Each patch added to the output texture is chosen by comparing the already known parts, depicted in figure 6.29 a), with all the possible positions in the input patch. The SSE criterion is then used to find the best matching area. That is, the patch size is determined by previously described DCT-descriptors.

Most of the time, the added patch does not match exactly the already synthesized texture, then a second step consists on cutting the patch with the texture using a graphcut like method described in [96]. Costs between overlapping patches is computed from the extended version in [96] using gradients. Figure 6.29 presents a typical case due to surrounding blocks and synthesis order.

6.3.6 Comparing and switching algorithms strategy

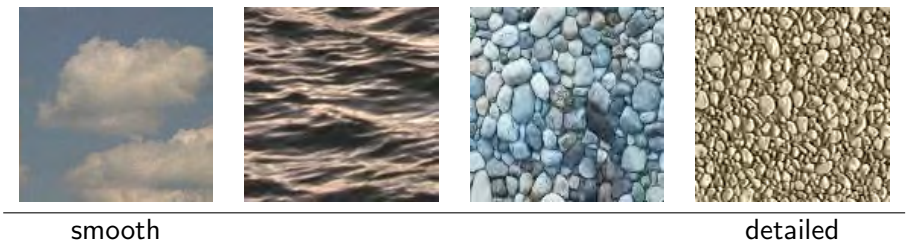


Figure 6.30: Texture spectrum for switching synthesizers

One has now to determine whether to use patch- or pixel-based technique depending on textures' characteristics. Then an experimental classification of textures in our context is presented in figure 6.30. Patch-based method provides good results with detailed textures but fails with smoother ones, creating edge artefacts. Reversely, pixel-based synthesis is well-adapted to smooth variations but fails when producing large detailed patterns. According to experimental results depicted in figures 6.31 and 6.32, patch-based method seems to give good results on grass and trodden earth but fails to synthesize water or clouds where pixel-based method produces better results.

(a) original frame (b) Missing textures (c) Pixel-based (d) Patch-based

Figure 6.31: Results on the *Coastguard* sequence

(a) original frame (b) Missing textures (c) Pixel-based (d) Patch-based

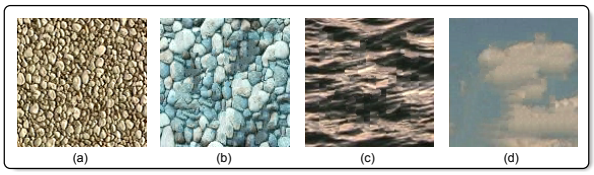
Figure 6.32: Results on the *Morocco Trial* sequence

Since the patch-based approach produces visible artifacts when failing, an a posteriori method is proposed. Three gradient-based criteria are proposed in order to quantify the amount of created artifacts. ΔGI denotes the gradient's differences at cut locations between synthesized and original images, while ΔGh and ΔGv represent horizontal and vertical mean Sobel gradients' differences, again between output and input textures. After experiments on a large set of textures, maximum thresholds $\{5; 5; 20\}$ have been fixed for ΔGh, ΔGv and ΔGI respectively to determine acceptable synthesis. GI characterizes the located seams to determine if the cuts are hidden enough, while Gh and Gv aim at pointing out such oriented edges, highly detectable by the Human Visual System, like in figure 6.33 c). As the texture is expected synthesizable by characterization step, the pixel-based synthesizer is chosen for high gradients.

The solution has been integrated to JM encoder [193]. Table 6.1 shows the coding performances obtained on *Container, Coastguard* et *Wool*. QP values {10,15,20,25,30,35,40} allow to measure how the removed regions impact the resulting bitrate.

6.3.7 Extension to video coding purposes

From this intra-frame compression framework using two kinds of texture synthesis, according to their favorite types of texture, an extension to image sequences has also been proposed in [151]. The idea



Texture	ΔGh	ΔGv	ΔGl
a	-0.3	-0.8	0.93
b	-3.5	-4.1	17.6
c	-12.3	11.1	28.7
d	-4.0	15.3	60.7

Figure 6.33: Resulting gradients for the 4 textures

QP	10	15	20	25	30	35	40
Sequences	Bit-rate saving (%)						
Coastguard CIF	21,8	22,4	20,7	20,3	19,7	19,5	18,9
Container CIF	21,7	19,3	14,4	10,6	7,1	4,5	2,7
Wool SD	14,4	13,9	11,7	10,3	8,8	7,5	7,1

Table 6.1: Resulting bit-rate saving with respect to the Quantization parameter QP

is to propose a synthesis solution that takes account of both spatial and temporal context. To this aim, both coding and synthesis modes should be carefully tuned so that to prevent from temporal divergence. To insure consistency between synthesized frames and the other ones, non synthesized frames, called reference frames, should be designed. Derived from GOP standard structures, we define GOPM (motion-oriented GOP) structures as a set of frames that are delimited by two reference frames.

The global framework is shown in figure 6.34. Reference images are classically encoded through MPEG-AVC scheme. For other frames, a temporal segmentation is first applied, so that to detect motion consistent regions. These regions will be encoded through adapted motion estimation/compensation methods. As for other regions, namely textured regions, they are processed by the switch pixel/patch based synthesis framework presented in 6.3.4. The resulting bistream then contains

- data related to macroblocks that are classically encoded,
- parameters of synthesis algorithms of each synthesized region,
- motion parameters for compensated regions (for affine motion, 6 parameters are required by region),
- information about location of synthesized or interpolated macroblocks.

Assessing the quality of decoded frames by such schemes, using objective methods, remains problematic. No satisfied solution has been until now found to assess synthesized textures within video compression context. Here, results on bit-rate savings are presented with the assumption of similar visual quality in order to tend to fair comparisons. To this aim, subjective tests have been lead to select videos. At comparable visual quality, up to 3% bit-rate is preserved, compared to H.264/AVC, on many SD and CIF sequences. As an example, figure 6.35 presents bit-rate savings according to 3 CIF sequences.

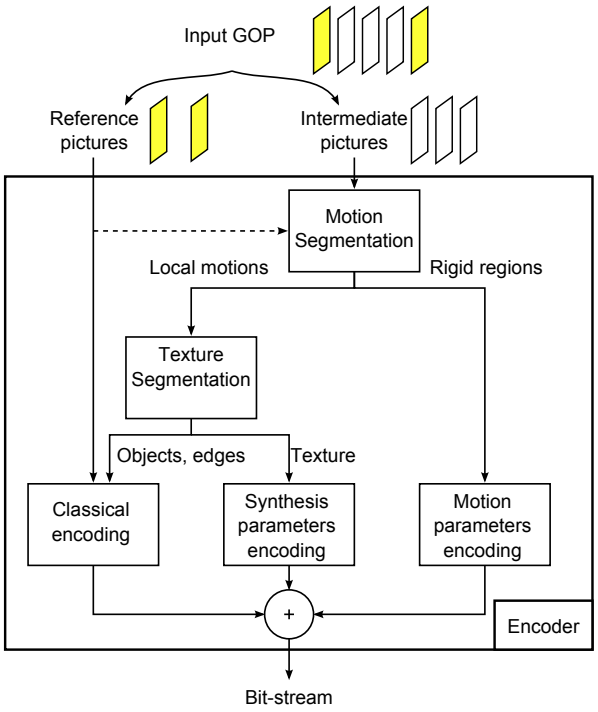


Figure 6.34: Global video encoding framework

6.4 Conclusion

A temporal consistent region representation has been first described. If the compression results are clearly not competitive with any video coder from the state-of-the-art, the major functionality, *i.e* the persistence of the regions throughout the sequence, remains of major interest. Based on a temporal RAG, homogeneous regions can be tracked.

With motion tubes, image sequences are interpreted as a set of patches of textures undergoing specific displacements and deformations along a given time interval. Indeed, each motion tube consists of an initial patch of texture, a lifespan during which the patch is available, and a set of warping operations describing its evolutions across time and space. In order to make their use easier, motion tubes were grouped into *families*: a reference image was partitioned into blocks, each of which initializing a motion tube. With motion tubes, the motion information now consists of a set of spatio-temporal trajectories along with local deformations. The temporal persistence of the texture is then naturally exhibited by the representation. Results show that *tubes* can effectively be used to represent image sequences, and coding improvements should be realized so that to be more competitive towards H.264/AVC standard.

As for the adaptive image synthesis tool for video compression, this approach is designed to be jointly used with current and future standard compression schemes. At encoder side, texture analysis includes segmentation and characterization tools, in order to localize candidate regions for synthesis. The corresponding areas are not encoded. The decoder fills them using texture synthesis. The remaining regions in images are classically encoded. They can potentially serve as input for texture synthesis. The chosen tools are developed and adapted with an eye to ensuring the coherency of

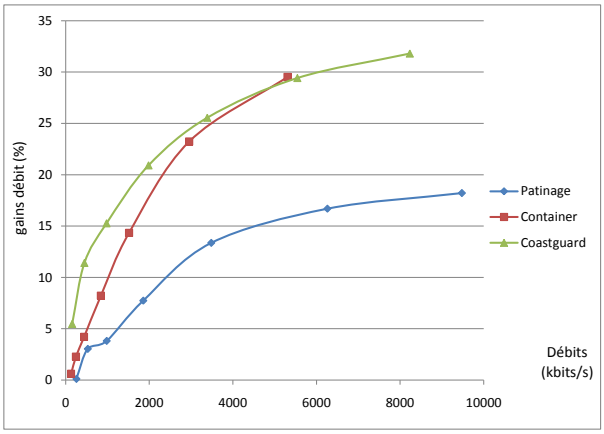


Figure 6.35: Bit-rate savings

the whole scheme. Thus, a texture characterization step provides required parameters to the texture synthesizer. Two texture synthesizers, including a pixel-based and a patch-based approach, are used on different types of texture, complementing each other. A first scheme is proposed for intra frame coding. Then, a temporal method is developed. The scheme is coupled with a motion estimator in order to segment coherent regions and to interpolate rigid motions using an affine model. Inter frame adapted synthesis is therefore used for non-rigid texture regions. Assessing the quality of decoded frames by such schemes, using objective methods, is problematic. Results on bit-rate savings are presented with the assumption of similar visual quality. At comparable visual quality, up to 33% bit-rate is preserved, compared to H.264/AVC, on many SD and CIF sequences.

Once again, if the coding issue was the final objective of these works, the resulting video representation allows advanced semantically based processes, such as patch or region tracking. When taking into account additional features, able to distinguish objects within a given scene, object tracking can be envisaged. In this way, the next chapter proposes some application contexts relative to robotic vision.

Chapter 7

Toward robotic vision for personal assistance living

Previous chapters have described joint video analysis and coding tools. If my first research objectives were related to compression, I am convinced that the proposed pseudo-semantic video representations have potential applications in other domains. Typically, geometrical cues, object detection and tracking can be deduced from advanced processes such as the ones previously presented.

Robotic vision shares common parts with video coding tools, especially 3D object tracking and semantical scene analysis. Through collaborative projects, that I manage and describe in this chapter, I operate a noteworthy change in my research projects. Together with embedded system concerns, I expect to design robotic vision systems dedicated to personal assistance living issues. Yao Zhigang, PhD student, has started his work which matches this context, whereas Rafiq Sekkal PhD's research topic moves slightly towards robotic vision.

This chapter then presents a roadmap that I intend to follow. General ideas and research perspectives are then exposed. In section 7.1, personal assistance living systems issues are described in the light of autonomy needs. Within this context, section 7.2 is dedicated to secured and autonomous navigation solution of electrical wheelchairs, whereas section 7.3 proposes to use 3D cameras to automatically detect unusual situations. Since these research axis are closely related to collaborative projects, these latter will be naturally described in corresponding sections.

7.1 Personal assistance living: towards higher autonomy

The loss of autonomy can be a consequence of ageing, evolutive or chronic diseases or even accident that induced permanent handicaps. Disabling chronic conditions have thus to be carefully considered in order to investigate technical and societal solutions able to preserve people autonomy.

When considering elderly people, their number and their proportion are expected to continuously increase. In France, demographic projections tend to show that the French population will stabilize at 64 million in 2050. By that time, seniors could account for 35 percent of the population. Since the Sixties, the health of elderly people in France has undergone significant changes. A transition has been observed, from acute illnesses to chronic or degenerative illnesses that lead to severe disabilities. While acute illness does not prevent seniors from travelling or having an active life,

the evolution up to disabling chronic diseases requires the adaptation and gradual introduction of activities, accommodation and nursing homes. This situation is not only relative to ageing and elderly people, but also concerns more generally disabled people.

In [58], it has been emphasized that the ability to freely move, whatever the destination and the timing are, remains a fundamental need for well-being and living well at home. For seniors, as soon as they loose their capacity to move by themselves, their mental condition is often affected, thus leading to additional handicaps. In [137], it has been shown that the notions of autonomy and integrity are strongly linked. In this way, it is obvious that creating tools towards autonomy and insuring secured moves are considered as a necessity.

As a consequence, personal assistance living issues are partially in relation with assistance robotic innovations. When considering robotic vision, performing automatic tasks with the help of cameras remains possible. The specificity of assistance robotic is that, as long as robotic systems are developed towards personal assistance, highly secure processes have to be designed.

The next section focuses then on a vision-based solution for wheelchair autonomous navigation.

7.2 Secured navigation of wheelchair solutions

7.2.1 APASH project

The APASH (Assistance au Pilotage pour l'Autonomie et la Sécurité des personnes Handicapées) project relies on a consortium composed of AdvanSEE (SME specialized in embedded vision systems), Ergovie (SME that sells and adapts wheelchairs) and three laboratories within INSA, namely IRISA, LGCGM (Sylvain Guégan) and IETR (Muriel Pressigout, Sylvain Haese, Luce Morin, François Pasteau). This project aims to develop technologies to assist disabled people and allow them to freely move indoor despite their handicap. This solution is particularly adapted to elderly people suffering from disorientation and spatio-temporal confusion.

Indeed, the use of electrical wheelchair requires three levels of sequential decisions:

- strategy level, *i.e.* decisions about trajectory for instance,
- tactical level, *i.e.* decisions about motion with smooth time constraints, such as slowing down, avoiding obstacles...
- operative level, *i.e.* decisions that require hard time constraints, such as avoiding imminent danger.

Moving with the help of a wheelchair remains then a quite complex task which requires both cognitive and sensory capacities. These capacities can be deeply altered, implying an increased difficulty to drive the wheelchair. In particular, going through doors, taking the elevator in a secure way without risking collision because of hazardous wheelchair motions remains a relevant issue.

Within this context, a first step of this research and development project consists of designing solutions to insure secured moves inside problematic areas. The main innovations brought by the APASH project rely then on the joint use of image information and heterogeneous multi-sensor data, in order to schedule vehicle trajectories. This would provide flexibility for an automatic navigation system and would ensure the immediate compatibility of the proposed system with existing electric wheelchairs.

Localization issues will be addressed through the use of robotic vision. Coupled with this localization process, navigation issues will require the fusion of data obtained from not only vision sensors but also from multiple other odometry sensors. The objective is to compute a motion controler for the electric wheelchair that takes account of the dynamic environment of hospitals.

The technical framework requires sensor networks: the resulting information redundancy ensures the necessary project reliability and safety, complying with the ISO26262 norm that focuses particularly on the transport of people. Performances will be evaluated in terms of patient acceptability and safety. The ergonomics of the wheelchair commands will be particularly studied so that to design appropriate and smart man-machine interface.

Further advanced researches are envisaged as soon as this first step of development will be available. Complete autonomous indoor and outdoor navigation will be the next social and technical challenge. Indeed, acquiring a high level of autonomy involves, in case of severe handicaps, to entirely assist people for their daily travels.

7.2.2 Technical issues

The technical objective of this work consists of developing a navigation system to ensure safe movement without collision of mobile robots in their dynamic environment. In order to improve the reactivity of mobile robots in their dynamic environment, we propose to take into account the full processing chain, *i.e.* from the sensor up to the model. To reach this objective, it will be necessary to jointly work on two aspects, namely the navigation system and the dynamic modeling of the mobile robot. The full model obtained by the interaction between these two tools will make possible to improve the control of the mobile robot. Zhigang Yao, a PhD student I co-supervise, has started his work relative to this framework in November 2011.

To obtain continuous indoor motion, vision based processes can be used. From a single embedded camera, obtained data have to be analyzed in order to be reactive enough in a dynamical environment. The main challenge of the system is to precisely determine the pose of the wheelchair. Related initialization and tracking features are still difficult problems especially in a dynamic environment [59]. Three types of approaches are then available: feature-based trackers relying on geometrical cues [35], template-matching approaches (*i.e.* based on texture recognition) [90], or hybrid solutions such as in [150]. Each of these solution addresses different application contexts as their robustness to occlusion may vary. The idea is then to use a hybrid-based solution to guarantee robust and real-time tracking. Markless solutions are here mandatory, as a complete and precise 3D model of the building will not necessary be available. In any case, additional techniques can be coupled in order to take account of aberrant estimated values and to improve robustness. Classically, M-estimators [35] or RANSAC [176] are used to this aim.

Mobile robot navigation will be then insured by visual servoing. Indeed, wheelchair oriented navigation issues are quite similar to the mobile robotic ones in terms of navigation strategies [157][120][36] and in terms of control [3].

As for it, the geometrical model of the wheelchair has to be elaborated. Coupled with both the kinematic and the dynamic models of the mobile robot, the optimal position of the sensors relatively to the wheelchair itself has to be found. The idea is then to identify all the possible motions of the wheelchair in order to precisely adapt the motion control. Deduced from these models, motions of the wheelchair should be as precise and smooth as possible in order to be safe for both the patient and the environment.

In addition, we have also to take into account the fact that the mobile robot will evolve in a dynamic environment. The task for visual control thus naturally varies according to the considered environment. For instance, the target (door, elevator,...) can be partially or totally occulted by a moving object thus requiring solution to estimate target position while moving the vehicle. In addition, for example when changing room, automatical definition of visual servoing reference will be a crucial issue.

Finally, the behavior of the vehicle has to be fully characterized, in particular when taking into account man-machine interactions. The control of this overall behavior is the key feature to guarantee security. Even if the wheelchair remains under control of the disabled people, the position and the velocity of the vehicle should match the environment constraints. As a consequence, the control unit will have to find a tradeoff between autonomous and remote control [187]. For instance, in case of obstacle or door detection, a solution consists of progressively substituting the human control for automatical control, as the obstacle is getting closer.

7.3 Fall detection through 3D cameras

Staying at home, for disabled or ageing people, has been proved to be a source of well-being. In-home care services can be then set up as soon as the concerned people agree. However, for family circle, or for people themselves, this matter of fact can be difficult to accept because of health risk. In particular, falls can rapidly become a case of emergency.

In the state-of-the-art, emergent technical solutions, relative to home-care support, tend to reassure people in the sense that they are able to raise alarm as soon as unusual situation is detected. To this aim, dedicated multiple camera-based frameworks have been developed [9]. In parallel, thanks to low-cost 3D cameras such as the Kinect, fall detection frameworks have been designed [163]. Meanwhile, the combination of these two methods has led to innovative solutions for gait analysis [8].

The project PATH4FAR (Posture Analysis and Tracking at Home for Fall AlaRm), led by NeoTec Vision (Rennes), aims at providing a complete framework for robust fall detection. From accurate analysis tools, together with dedicated embedded system, we intend to verify the ability of the framework to provide "true" alarms. This project, until now, is not yet financed. It will be realized with the help of the Hospital of Port Louis.

The idea is here to adapt robust 3D tracking solutions to depth signals in order to precisely localize people as well as their positions in the room. Multiple 3D cameras will be used to this aim. In particular, challenges lie in the fact that the resolution of depth signals is not as accurate as a monocular camera.

7.4 Long term objectives

A transversal topic of these projects relies on tracking issues. From compression up to robotic vision, tracking textures, objects or geometrical cues are common issues. In Lagadic team, some previous related works have already been realized especially towards markerless and robust tracking systems. Yet some issues, such as tracking initialization or real-time object detection, have not been completely solved. PhD work of Rafiq Sekkal will address these sensitive issues through joint object extraction

and tracking. In particular, his previous work, based on multiresolution segmentation (see section 5.3), could be seen as an answer to the visual servoing initialization issue.

As for the autonomous navigation, an approach based on both visual servoing laws and appearance-based navigation from an image database has been designed for urban vehicle navigation purposes (ANR projects Predit MobiVIP and CityVIP). The idea is to verify if this technique can be applied within the APASH project context. APASH related application implies a maximal safety while moving in a very constraint environment (in hospital or at home, obstacle and stairs avoidance...). In order to enhance the wheelchair motion safety, additional information from heterogenous sensors should be provided. In particular, radio technologies such as Ultra Wide Band (UWB) systems can provide indoor precise localization [68][60][44]. In collaboration with Sylvain Haese (IETR - CPR group), we intend to combine UWB and 3D vision to improve localization precision, especially in high danger areas such as stairs. Dedicated sensor fusion process would thus be studied.

In a nutshell, these long term objectives naturally confirm the fact that my research areas are now oriented towards robotic vision issues. In conjunction with embedded systems concern, applications that are envisaged are mainly focused on personal assistance living.

List of Figures

List of Tables

Bibliography

[1] A. M. T. A. Leontaris. Rate control reorganization in the joint model soft reference. Technical report, Joint Video Team, doc. JVT-W042, 2007.

[2] A. Albanese, J. Blömer, J. Edmonds, M. Luby, and M. Sudan. Priority Encoding Transmission. *IEEE Transaction on Information Theory*, 42(6):1737–1744, 1996.

[3] G. Allibert, E. Courtial, and F. Chaumette. Predictive control for constrained image-based visual servoing. *Robotics, IEEE Transactions on*, 26(5):933 –939, oct. 2010.

[4] Y. Altunbasak and N. Kamaci. An analysis of the DCT coefficient distribution with the h.264 video coder. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, volume 3, pages 177–80. IEEE, May 2004.

[5] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2294–2301, 2009.

[6] M. Ashikhmin. Synthesizing natural textures. In *Proceedings of ACM Symposium on Interactive 3D Graphics*, pages 217–226, 2001.

[7] N. Asuni and A. Giachetti. Accuracy improvements and artifacts removal in edge based image interpolation. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 58–65, 2008.

[8] E. Auvinet, F. Multon, and J. Meunier. Contactless Abnormal Gait Detection, Gait analysis with multiple depth cameras. In *EMBC*, Boston, États-Unis, 2011. IEEE, IEEE Press.

[9] E. Auvinet, F. Multon, A. Saint-Arnaud, J. Rousseau, and J. Meunier. Fall detection with multiple cameras: An occlusion-resistant method based on 3-d silhouette vertical distribution. *Information Technology in Biomedicine, IEEE Transactions on*, 15(2):290 –300, march 2011.

[10] W. A. B. Perceptual-components architecture for digital video. *Journal of the Optical Society of America. A: Optics and image science, volume 7, issue 10*, pages 1943–1954, Oct. 1990.

[11] M. Babel, L. Bédat, O. Déforges, and J. Motsch. Context-Based Scalable Coding and Representation of High Resolution Art Pictures for Remote Data Access. In *Proc. of the IEEE International Conference on Multimedia and Expo, ICME'07*, pages 460–463, July 2007. Projet ANR TSAR.

[12] M. Babel and O. Déforges. wg1n4870 Response to call for AIC technologies and evaluation methodologies. Technical report, ISO/ITU JPEG commitee, San Francisco, USA, January 2009.

[13] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5327 Answer for the AIC call - Medical Images database . Technical report, ISO/ITU JPEG commitee, Boston, USA, March 2010.

[14] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5545 LAR Codec reference software overview - Core Experiment 1.0 . Technical report, ISO/ITU JPEG commitee, Guangzhou, China, August 2010.

[15] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5630 LAR Codec reference software overview - Intermediate Core Experiment 1.0. Technical report, ISO/ITU JPEG commitee, December 2010.

[16] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5631 LAR Codec reference software - Core Experiment 2.0 . Technical report, ISO/ITU JPEG commitee, January 2011.

[17] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5804 LAR Codec reference software - Core Experiment 3.0. Technical report, ISO/ITU JPEG commitee, Berlin, July 2011.

[18] M. Babel, O. Déforges, L. Bédat, C. Strauss, and F. Pasteau. wg1n5812 LAR Objective Quality Comparison. Technical report, ISO/ITU JPEG commitee, Berlin, July 2011.

[19] M. Babel, O. Déforges, L. Bédat, C. Strauss, F. Pasteau, and J. Motsch. WG1N5315 - Response to Call for AIC evaluation methodologies and compression technologies for medical images: LAR Codec. Technical report, ISO/ITU JPEG commitee, Boston, USA, March 2010.

[20] M. Babel, O. Déforges, L. Bédat, C. Strauss, F. Pasteau, and J. Motsch. wg1n5330 1003 Response - CFP AIC Medical images - IETR . Technical report, ISO/ITU JPEG commitee, Boston, USA, March 2010.

[21] M. Babel, O. Déforges, and J. Ronsin. Interleaved S+P Pyramidal Decomposition with Refined Prediction Model. In *IEEE International Conference on Image Processing, ICIP'05*, pages 750–753. IEEE, Sept. 2005.

[22] M. Babel, B. Parrein, O. Déforges, N. Normand, J.-P. Guédon, and V. Coat. Joint source-channel coding: secured and progressive transmission of compressed medical images on the Internet. *Computerized Medical Imaging and Graphics*, 32(4):258–269, June 2008.

[23] M. Babel, F. Pasteau, C. Strauss, M. Pelcat, L. Bédat, B. Médéric, and O. Déforges. Preserving data integrity of encoded medical images: the LAR compression framework. In R. Kountchev and K. Nakamatsu, editors, *Advances in reasoning-based image processing, analysis and intelligent systems: Conventional and intelligent paradigms*, Intelligent Systems Reference Library, pages 1–36. Springer, 2011.

[24] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3/4):313 – 336, 1996.

[25] C. Bergeron, B. Gadat, C. Poulliat, and D. Nicholson. Extrinsic distortion based Source-Channel allocation for Wireless JPEG2000 transcoding systems. In *IEEE International Conference on Image Processing*, Hong-Kong, September 2010.

[26] S. S. Bhattacharyya, J. Eker, J. Janneck, C. Lucarz, M. Mattavelli, and M. Raulet. Overview of the MPEG Reconfigurable Video Coding Framework. *Journal of Signal Processing Systems*, 63(2):251–263, 2009.

[27] H. Boeglen. IT++ library for numerical communications simulations. *http://herve.boeglen.free.fr/itpp_windows/*, 2007.

[28] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *IEEE 12th International Conference on Computer Vision*, pages 833–840, oct 2009.

[29] N. Cammas and S. Pateux. Fine grain scalable video coding using 3d wavelets and active meshes. In *SPIE Visual Communications and Image Processing*, Jan 2003.

[30] W. Carey, D. Chuang, and S. Hemami. Regularity-preserving image interpolation. *IEEE Transactions on Image Processing*, 8(9):1293–1297, Sept. 1999.

[31] X. Chen, H. Ou, X. Luo, M. Chen, Y. Zhang, K. Hao, and S. Mi. The Progress of University Digital Museum Grid. In *Proc. IEEE International Conference on Grid and Cooperative Computing Workshops, GCCW'06*, 2006.

[32] W.-C. S. Chi-Shing Wong. Further Improved Edge-directed Interpolation and Fast EDI for SDTV TO HDTV Conversion. In *EUSIPCO 2010*, Aalborg, Denmark, August 2010.

[33] B.-D. Choi, J.-W. Han, S.-W. Jung, J.-H. Nam, and S.-J. Ko. Overlapped Block Motion Compensation based on irregular grid. In *Proceedings of International Conference on Image Processing (ICIP'06)*, pages 1085–1088, 2006.

[34] CIE. CIE publication 116-1995 "Industrial Colour-Difference evaluation", 1995.

[35] A. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *Visualization and Computer Graphics, IEEE Transactions on*, 12(4):615 –628, july-aug. 2006.

[36] J. Courbon, Y. Mezouar, and P. Martinet. Indoor navigation of a non-holonomic mobile robot using a visual memory. *Auton. Robots*, 25:253–266, October 2008.

[37] J. Cristy, A. Thyssen, and F. Weinhaus. ImageMagick. http://www.imagemagick.org.

[38] O. Deforges. *Codage d'images par la méthode LAR et Méthodologie Adéquation Algorithme Architecture : de la définition des algorithmes de compression ou prototypage rapide sur architectures parallèles hétérogènes*. Habilitation thesis, Université de Rennes 1, November 2004.

[39] O. Déforges and M. Babel. LAR method: from algorithm to synthesis for an embedded low complexity image coder. In *3rd International Design and Test Workshop - IDT'08*, December 2008.

[40] O. Déforges, M. Babel, L. Bédat, and J. Ronsin. Color LAR codec: a color image representation and compression scheme based on local resolution adjustment and self-extracting region representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(8):974–987, 2007.

[41] O. Déforges, M. Babel, and J. Motsch. The RWHT+P for an improved lossless multiresolution coding. In *EUSIPCO proceedings*, pages 1–5. EUSIPCO, 2006.

[42] O. Deforges and J. Ronsin. Locally adaptive resolution method for progressive still image coding. In *Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on*, volume 2, pages 825 –829 vol.2, 1999.

[43] B. C. Dharaa and B. Chanda. Color image compression based on block truncation coding using pattern fitting principle. In *Pattern Recognition*, volume 40, pages 2408–2417, September 2007.

[44] A. Dilshab and S. Chauhan. Latest advancements in mobile robot localization in manufacturing environment. In *Second International Conference on Business and Technology*, 2010.

[45] J. Dong and N. Ling. On model parameter estimation for h.264/avc rate control. In *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, pages 289 –292, may 2007.

[46] C. E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.

[47] F. Dufaux and D. Nicholson. JWL: JPEG 2000 for wireless applications. In *SPIE Proc. Applications of Digital Image Processing XXVII*, volume 5558, pages 309–318, 2004.

[48] A. Dumitras, B. Haskell, A. Inc, and C. Cupertino. A texture replacement method at the encoder for bit-rate reduction of compressed video. *IEEE TCSVT*, 13(2):163–175, 2003.

[49] P. T. Ebrahimi. wg1n5578 WebP vs. Standard Image Coding. Technical report, ISO/ITU JPEG commitee, ICT-Link, October 2010.

[50] J. Editors. JPEG 2000 image coding system - Part 11: Wireless JPEG2000 Committee Draft. in ISO/IEC CD 15444-11 / ITU-T SG8, 2004.

[51] J. Eker and J. W. Janneck. Cal language report. Technical Report ERL Technical Memorandum No. UCB/ERL M03/48, University of California at Berkeley, December 2003.

[52] P. Elias. Coding for Noisy Channels. *Convention Record*, 4:37–49, 1955.

[53] M. Fairchild. *Color Appearance Models*. John Wiley & Sons, Ltd, England, 2005.

[54] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[55] E. Flécher, M. Babel, O. Déforges, and V. Coat. LAR Video: Hierarchical Representation for Low Bit-Rate Color Image Sequence Coding. In *Proc. of Picture Coding Symposium (PCS'07)*, page paper 1175, Lisboa, Portugal, Nov. 2007.

[56] E. Flécher, M. Raulet, G. Roquier, M. Babel, and O. Déforges. Framework For Efficient Cosimulation And Fast Prototyping on Multi-Components With AAA Methodology: LAR Codec Study Case. In *Proc. of the 15th European Signal Processing Conference (Eusipco 2007)*, pages 1667–1671, Poznań, Poland, September 2007.

[57] C. Fonteneau, J. Motsch, M. Babel, and O. Déforges. A Hierarchical Selective Encryption Technique in a Scalable Image Codec. In *Proc. of International Conference in Communications*, June 2008.

[58] A. Franco. Rapport de la mission "Vivre chez soi". Report to mrs nora berra, secretary of state of seniors, French ministry of Work, Health and Solidarity, June 2010.

[59] P. Fua and V. Lepetit. *Emerging Technologies of Augmented Reality, Interfaces and Design*, chapter Vision Based 3D Tracking and Pose Estimation for Mixed Reality, pages 1–22. Idea Group Publishing., 2007.

[60] S. Gezici, Z. Tian, G. Giannakis, H. Kobayashi, A. Molisch, H. Poor, and Z. Sahinoglu. Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. *Signal Processing Magazine, IEEE*, 22(4):70 – 84, july 2005.

[61] A. Giachetti and N. Asuni. Real time artifact-free image upscaling. *Image Processing, IEEE Transactions on*, page 1, 2011.

[62] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009.

[63] S. Golomb. Run-length encodings (Corresp.). *IEEE Transactions on Information Theory*, 12(3):399–401, July 1966.

[64] google.com. Webm: an open web media project. http://www.webmproject.org/tools/vp8-sdk/, 2011.

[65] google.com. Webp: A new image format for the web. http://code.google.com/speed/webp/, 2011.

[66] M. J. Gormish. Lossless and nearly lossless compression for high-quality images. In *Proceedings of SPIE*, pages 62–70, San Jose, CA, USA, 1997.

[67] T. Grandpierre and Y. Sorel. From algorithm and architecture specifications to automatic generation of distributed real-time executives: a seamless flow of graphs transformations. In *First ACM and IEEE International Conference on Formal Methods and Models for Co-Design*, Mont Saint-Michel, France, June 2003.

[68] R. Grossmann, J. Blumenthal, F. Golatowski, and D. Timmermann. Localization in zigbee-based sensor networks. In *1st European Zigbee Developer's Conference (EuZDC)*, München-Dornach, Deutschland, 2007.

[69] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions $*$. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1030–1037, 2009.

[70] H264 MPEG-4 10 AVC. *Joint Committee Draft (CD)*. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEGn 3rd Meeting: Fairfax, Virginia, USA, May 2002.

[71] W. Hamidouche, C. Olivier, M. Babel, O. Déforges, H. Boeglen, and P. Lorenz. LAR Image transmission over fading channels: a hierarchical protection solution. In *Proc. of The Second International Conference on Communication Theory, Reliability, and Quality of Service*, pages 1–4, Colmar France, 07 2009.

[72] Z. He and T. Chen. Linear rate control for JVT video coding. *Information Tech.: Research and Education, Int. Conf. on*, pages 65–68, 2003.

[73] Z. He and S. Mitra. ρ-domain bit allocation and rate control for real time video coding. *Image Processing, Int. Conf. on*, 3:546–549, 2001.

[74] D. A. Huffman. A method for the construction of Minimum-Redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, Sept. 1952.

[75] R. S. Hunter. Photoelectric color difference meter. *Journal of the Optical Society of America*, 48(12):985, Dec. 1958.

[76] P. Ishwar and P. Moulin. Switched control grid interpolation for motion compensated video coding. In *Proceedings of International Conference on Image Processing (ICIP'97)*, volume 3, pages 650–653, 1997.

[77] P. Ishwar and P. Moulin. On spatial adaptation of motion field smoothness in video coding. *IEEE Transactions on circuits and systems for video technology*, 10(6):980–989, Sep 2000.

[78] ISO/IEC 10918-1. *Information technology - Digital compression and coding of continuous-tone still images - Requirements and guidelines*, Mars 2009.

[79] ISO/IEC 14495-1. *Information technology - Lossless and near-lossless compression of continuous-tone still images - Baseline*, Mars 2009.

[80] ISO/IEC 15444-1. *JPEG 2000 image coding system Part 1*, May 2000.

[81] ISO/IEC 15444-1. *(JPEG-1)-based still-image coding using an alternative arithmetic coder* , May 2006.

[82] ISO/IEC 15444-1. *JPEG 2000 image coding system: Wireless* , May 2006.

[83] ISO/IEC 15444-1. *JPEG XR image coding system: Image coding specification* , Mars 2009.

[84] ISO/IEC JTC1/SC29/WG11 MPEG2007/N9189. Svc verification test plan, version 1. Technical report, Joint Video Team, 2007.

[85] ITU-T. Recommendation E.800: QoS Terms and Definitions related to Quality of Service and Network Performance including dependability. Technical report, International Telecommunication Union, August 1994.

[86] ITU-T. Recommendation G.1080: Quality of Experience requirements for IPTV services. Technical report, International Telecommunication Union, May 2008.

[87] F. Jamil, R. Porle, A. Chekima, R. Lee, H. Ali, and S. Rasat. Preliminary study of block matching algorithm (bma) for video coding. In *Mechatronics (ICOM), 2011 4th International Conference On*, pages 1 –5, may 2011.

[88] K. Jerbi, M. Wipliez, M. Raulet, O. Déforges, M. Babel, and M. Abid. Automatic Method For Efficient Hardware Implementation From RVC-CAL Dataflow: A LAR Coder baseline Case Study. *"Section E: Consumer Electronics" of the Journal of Convergence*, pages 85–92, Dec. 2010.

[89] K. Jerbi, M. Wipliez, M. Raulet, O. Déforges, M. Babel, and M. Abid. Fast Hardware implementation of an Hadamard Transform Using RVC-CAL Dataflow Programming. In *Proc. of 5th IEEE International Conference on Embedded and Multimedia Computing 5th IEEE International Conference on Embedded and Multimedia Computing*, pages 1–5, Philippines, 08 2010.

[90] F. Jurie and M. Dhome. Real time robust template matching. In *in British Machine Vision Conference 2002*, pages 123–131, 2002.

[91] D. Kelkar and S. Gupta. Improved quadtree method for split merge image segmentation. In *IEEE International Conference on Emerging Trends in Engineering and Technology*, volume 0, pages 44–47, Los Alamitos, USA, 2008.

[92] W. Kropatsch, Y. Haxhimusa, and A. Ion. Multiresolution image segmentations in graph pyramids. In A. Kandel, H. Bunke, and M. Last, editors, *Applied Graph Theory in Computer Vision and Pattern Recognition*, volume 52 of *Studies in Computational Intelligence*, pages 3–41. Springer Berlin / Heidelberg, 2007.

[93] W. Kropatsch, Y. Haxhimusa, and A. Ion. Multiresolution image segmentations in graph pyramids. In A. Kandel, H. Bunke, and M. Last, editors, *Applied Graph Theory in Computer Vision and Pattern Recognition*, volume 52 of *Studies in Computational Intelligence*, pages 3–41. Springer Berlin / Heidelberg, 2007.

[94] W. G. Kropatsch, Y. Haxhimusa, Z. Pizlo, and G. Langs. Vision pyramids that do not grow too high. *Pattern Recogn. Lett.*, 26:319–337, February 2005.

[95] V. Kwatra, I. E. Aaron, and B. N. Kwatra. Texture Optimization for Example-based Synthesis. In *Proceedings of ACM SIGGRAPH*, pages 795 – 802, 2005.

[96] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: image and video synthesis using graph cuts. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, pages 277–286, 2003.

[97] E. Y. Lam and J. W. Goodman. A mathematical analysis of the DCT coefficient distributions for images. *IEEE Transactions on Image Processing*, 9(10):1661–1666, Oct. 2000.

[98] C. Larabi. wg1n4805 Call for Advanced Image Coding and evaluation methodologies (AIC). Technical report, ISO/ITU JPEG commitee, University of Poitiers, October 2008.

[99] C. Larabi. wg1n5491 AIC core experiment on evaluation of LAR proposal (CO-LAR-01). Technical report, ISO/ITU JPEG commitee, University of Poitiers, July 2010.

[100] C. Larabi. wg1n5596 AIC core experiment on evaluation of LAR proposal (CO-LAR-02). Technical report, ISO/ITU JPEG commitee, University of Poitiers, October 2010.

[101] C. Larabi. wg1n5682 Subjective evaluation results for CO-LAR-02. Technical report, ISO/ITU JPEG commitee, University of Poitiers, February 2011.

[102] C. Larabi. wg1n5712 AIC core experiment on performance evaluation and functionality analysis. Technical report, ISO/ITU JPEG commitee, University of Poitiers, February 2011.

[103] B. Le Guen. *Adaptation du contenu spatio-temporel des images pour un codage par ondelettes*. PhD thesis, Université de Rennes I, Rennes, France, Feb 2008.

[104] W. S. Lee. Edge adaptive prediction for lossless image coding. In *Proc. Data Compression Conference*, pages 483–490, 1999.

[105] X. Li, N. Oertel, A. Hutter, and A. Kaup. Laplace distribution based lagrangian rate distortion optimization for hybrid video coding. *IEEE Trans. Cir. and Sys. for Video Technol.*, 19:193–205, February 2009.

[106] X. Li and M. T. Orchard. New Edge-Directed interpolation. *IEEE Transactions on Image Processing*, 10:1521–1527, 2001.

[107] Z. Li, F. Pan, K. Lim, G. Feng, and X. Lin. Adaptive basic unit layer rate control for JVT. Technical report, Joint Video Team, doc. JVT-G012, 2003.

[108] L.-J. Lin and A. Ortega. Bit-rate control using piecewise approximated rate-distortion characteristics. *IEEE Trans. Circuits Syst. Video Tech.*, 8:446–459, 1998.

[109] D. Liu, X. Sun, F. Wu, S. Li, and Y.-Q. Zhang. Image compression with edge-based inpainting. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(10):1273 –1287, Oct. 2007.

[110] Y. Liu, Z. G. Li, and Y. C. Soh. Rate control of h.264/AVC scalable extension. *Circuits and Systems for Video Tech., IEEE Trans. on*, 18(1):116–121, 2008.

[111] Y. Liu and Y. Tsin. The promise and the perils of near-regular texture. *International Journal of Computer Vision*, 62:1–2, 2002.

[112] Z. Liu, L. Karam, and A. Watson. JPEG2000 encoding with perceptual distortion control. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I–637–40 vol.1, 2003.

[113] B. Ltd. PhotoZoomPro3. http://www.benvista.com/photozoompro.

[114] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, 2007.

[115] S. W. Ma, W. Gao, Y. Lu, and H. Q. Lu. Proposed draft description of rate control on JVT standard. Technical report, Joint Video Team, doc. JVT-F086, 2002.

[116] G. F. MacDonald. Digital visionary. *Museum News*, 79:34–41, March/April 2000.

[117] M. Maire, P. Arbelaez, C. C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[118] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

[119] H. Malvar and G. Sullivan. Ycocg-r: A color space with rgb reversibility and low dynamic range. In *Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG*, 22-24 July 2003.

[120] N. Mansard and F. Chaumette. Task sequencing for high-level sensor-based control. *Robotics, IEEE Transactions on*, 23(1):60 –72, feb. 2007.

[121] D. Marpe, H. Schwarz, S. Bosse, B. Bross, P. Helle, T. Hinz, H. Kirchhoffer, H. Lakshman, T. Nguyen, S. Oudin, M. Siekmann, K. Sühring, M. Winken, and T. Wiegand. Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding. *IEEE Trans. on Circuits and Systems for Video Technology*, pages 1676–1687, 2010.

[122] F. Marqués, M. Pardàs, and R. Morros. Object matching based on partition information. In *ICIP (2)*, pages 829–832, 2002.

[123] F. Marques, P. Salembier, M. Pardas, J. Morros, I. Corset, S. Jeannin, B. Marcotegui, and F. Meyer. A segmentation-based coding system allowing manipulation of objects (sesame). In *IEEE International Conference of Image Processing*, pages III: 679–682, 1996.

[124] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[125] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:530–549, May 2004.

[126] C. Mehlführer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp. Simulating the Long Term Evolution Physical Layer. In *Proc. of the 17th European Signal Processing Conference*, August 2009.

[127] M. Mignotte. Mds-based multiresolution nonlinear dimensionality reduction model for color image segmentation. *Neural Networks, IEEE Transactions on*, 22(3):447 –460, march 2011.

[128] A. Mohr, E. A. Riskin, and R. E. Ladner. Unequal Loss Protection : Graceful degradation of image quality over packet erasure channels through forward error correction. *Journal on Selected Areas in Communications*, 18(6):819–828, June 2000.

[129] L. Morin. Modélisation 3d pour la communication vidéo. Habilitation à Diriger des Recherches - IRISA-IFSIC, Université de Rennes 1, France, may 2006.

[130] J. Morros and F. Marques. A proposal for dependent optimization in scalable region-based coding systems. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 4, pages 295–299, October 1999.

[131] J. Motsch, M. Babel, and O. Déforges. Joint Lossless Coding and Reversible Data Embedding in a Multiresolution Still Image Coder. In *Proc. of European Signal Processing Conference, EUSIPCO*, pages 1–4, Glasgow UK, 08 2009.

[132] J. Motsch, O. Déforges, and M. Babel. Embedding Multilevel Image Encryption in the LAR Codec. In *IEEE Communications International Conference 06*, Bucharest, Romania, Juin 2006.

[133] P. Ndjiki-Nya, D. Doshkov, and M. Koppel. Optimization of video synthesis by means of cost-guided multimodal photometric correction. In *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, pages 111 –116, sept. 2009.

[134] P. Ndjiki-Nya, T. Hinz, and T. Wiegand. Generic and robust video coding with texture analysis and synthesis. In *2007 IEEE ICME*, pages 1447–1450, 2007.

[135] V. A. Nguyen, Z. Chen, and Y.-P. Tan. Video enhancement from multiple compressed copies in transform domain. *J. Image Video Process.*, 2010:5:1–5:18, January 2010.

[136] S. Nogaki and M. Ohta. An overlapped block motion compensation for high quality motion picture coding. In *Proceedings of IEEE International Symposium on Circuits and Systems*, volume 5, pages 427–440, Apr 1992.

[137] L. Nordenfelt. *Dignity in Care for Older People*. Wiley-Blackwell, 2009.

[138] T. Ochotta and D. Saupe. Edge-based partition coding for fractal image compression. *The Arabian Journal for Science and Engineering, Special Issue on Fractal and Wavelet Methods*, 29(2C):63–83, December 2004.

[139] M. S. Ogawa. wg1n5583 Compare experiment of execution speed JPEG/JPEG2000/JPEGXR. Technical report, ISO/ITU JPEG commitee, ICT-Link, October 2010.

[140] M. T. Orchard and G. J. Sullivan. Overlapped Block Motion Compensation: an estimation theoretic approach. *IEEE Transactions on Image Processing*, 3(5):693–699, Sep 1994.

[141] C. Pang, O. C. Au, J. Dai, F. Zou, W. Yang, and M. Yang. Laplacian mixture model(lmm) based frame-layer rate control method for h.264/avc high-definition video coding. In *ICME*, pages 25–29, 2010.

[142] F. Pasteau. *Statistical study of a predictive codec : a LAR-based robust and flexible framework*. PhD thesis, INSA Rennes, November 2011.

[143] F. Pasteau, M. Babel, O. Déforges, and L. Bédat. Interleaved S+P Scalable Coding with Inter-Coefficient Classification Methods. In *Proc. of the EUSIPCO'08*, pages 1–5, Lausanne, Suisse, Aug. 2008.

[144] F. Pasteau, C. Strauss, M. Babel, O. Déforges, and L. Bédat. Improved colour decorrelation for lossless colour image compression using the LAR codec. In *Proceedings of EUSIPCO'09*, pages 1–4, Glasgow, Royaume-Uni, Aug. 2009.

[145] F. Pasteau, C. Strauss, M. Babel, O. Déforges, and L. Bédat. Adaptive colour decorrelation for predictive image codecs. In *Proc. of EUSIPCO 2011*, pages 1–5, Barcelona, Espagne, Aug. 2011.

[146] B. Pesquet-Popescu. *Scalabilité et robustesse en codage vidéo*. Habilitation thesis, Université de Nice Sophia-Antipolis, October 2005.

[147] Y. Pitrey. *Stratégies d'encodage pour codeur vidéo scalable (Coding strategies for scalable video coder)*. PhD thesis, INSA Rennes, September 2009.

[148] Y. Pitrey, M. Babel, O. Déforges, and J. Vieron. ρ-domain based rate control scheme for spatial, temporal and quality scalable video coding. *SPIE Electronic Imaging (VCIP)*, 2008.

[149] D. Pitzalis, R. Pillay, and C. Lahanier. A new Concept in high Resolution Internet Image Browsing. In *10th International Conference on Electronic Publishing (ELPUB)*, June 2006.

[150] M. Pressigout and E. Marchand. Real-time 3d model-based tracking: Combining edge and texture information. In *IEEE ICRA06*, pages 2726–2731, 2006.

[151] F. Racapé. *Mise en Oeuvre de Techniques d'Analyse/Synthèse de Texture dans un Schéma de Compression Vidéo*. PhD thesis, INSA Rennes, November 2011.

[152] F. Racapé, S. Lefort, E. Francois, M. Babel, O. Déforges, and Racapé. Adaptive pixel/patch-based synthesis for texture compression. In *Proc. of ICIP 2011*, pages 1–4, Brussels, Belgique, Sept. 2011.

[153] F. Racapé, S. Lefort, D. Thoreau, M. Babel, and O. Déforges. Characterization and adaptive texture synthesis-based compression scheme. In *In Proc. of EUSIPCO 2011*, pages 1–5, Espagne, Aug. 2011.

[154] M. Raulet, M. Babel, J.-F. Nezan, O. Déforges, and Y. Sorel. Automatic Coarse Grain Partitioning and Automatic Code Generation for Heterogeneous Architectures. In *IEEE Workshop on Signal Processing Systems (SIPS'03)*, pages 27–29, Seoul, Korea, August 2003.

[155] I. Reed and G. Solomon. Polynomial Codes Over Certain Finite Fields. *Journal of the Society of Industrial and Applied Mathematics (SIAM)*, 2:300–304, June 1960.

[156] J. Reichel, H. Schwarz, and M. Wien. Scalable video coding - joint draft 4. Technical report, Joint Video Team, doc. JVT-Q201, 2005.

[157] A. Remazeilles and F. Chaumette. Image-based robot navigation from an image memory. *Robot. Auton. Syst.*, 55:345–356, April 2007.

[158] J. Ribas-Corbera and S. Lei. Rate control in DCT video coding for low-delay communications. *Circuits and Systems for Video Tech., IEEE Trans. on*, 9(1):172–185, Feb 1999.

[159] D. T. Richter. wg1n4558 JPEG-XR core experiment results with objectiv metrics. Technical report, ISO/ITU JPEG commitee, University of Stuttgart, April 2008.

[160] T. Richter. wg1n5681 AIC core experiment 2. Technical report, ISO/ITU JPEG commitee, Universität Stuttgart, February 2011.

[161] J. Rissanen and G. G. Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23(2):149–162, Mar. 1979.

[162] M. Robins. Delivering optimal Quality of Experience (QoE) for IPTV success. Whitepaper, Spirent Communications, February 2006.

[163] C. Rougier, E. Auvinet, J. Rousseau, M. Mignotte, and J. Meunier. Fall detection from depth map video sequences. In B. Abdulrazak, S. Giroux, B. Bouchard, H. Pigot, and M. Mokhtari, editors, *ICOST*, volume 6719 of *Lecture Notes in Computer Science*, pages 121–128. Springer, 2011.

[164] A. Roussos and P. Maragos. Reversible interpolation of vectorial images by an anisotropic Diffusion-Projection PDE. *International Journal of Computer Vision*, 84:130–145, 2009. 10.1007/s11263-008-0132-x.

[165] S. A. Ruzinski. SAR Image Processor. http://www.general-cathexis.com/.

[166] M. Sabha, P. Peers, and P. Dutre. Texture synthesis using exact neighborhood matching. *Computer Graphics Forum*, 26(2):131–142, 2007.

[167] K. Sakamura. Digital Museum - Distributed Museum Concept for the 21st Century, 2000.

[168] P. Salembier, F. Marques, M. Pardas, J. Morros, I. Corset, S. Jeannin, L. Bouchard, F. Meyer, and B. Marcotegui. Segmentation-based video coding system allowing the manipulation of objects. *Circuits and Systems for Video Technology, IEEE Transactions on*, 7(1):60 –74, feb 1997.

[169] P. Schelkens, A. Skodras, and T. Ebrahimi. *The JPEG 2000 Suite*. John Wiley and Sons, Oct. 2009.

[170] H. Schwarz, D. Marpe, and T. Wiegand. Hierarchical b pictures. Technical report, Joint Video Team, doc. JVT-P014, 2002.

[171] H. Schwarz, D. Marpe, and T. Wiegand. Further results on constrained inter-layer prediction. Technical report, Joint Video Team, doc. JVT-O074, 2005.

[172] N. Sebe, M. S. Lew, and D. P. Huijsmans. Toward improved ranking metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1132–1143, 2000.

[173] C. Shaffer and H. Samet. Optimal Quadtree Construction Algorithms. *Computer Vision, Graphics, Image processing*, 37(3):402–419, March 1987.

[174] G. Sharma, W. Wu, and E. N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30:21–30, 2005.

[175] I. Shin, Y. Lee, and H. Park. Rate control using linear rate-ρ model for h.264. *Signal Processing - Image Communication*, 4:341–352, 2004.

[176] G. Simon and M.-O. Berger. Pose estimation for planar structures. *Computer Graphics and Applications, IEEE*, 22(6):46 – 53, nov/dec 2002.

[177] F. D. Simone, D. Ticca, F. Dufaux, M. Ansorge, and T. Ebrahimi. A comparative study of color image compression standards using perceptually driven quality metrics. In *Proceedings of SPIE*, pages 70730Z–11, San Diego, CA, USA, 2008.

[178] F. Smach, C. Lemaître, J. Gauthier, J. Miteran, and M. Atri. Generalized Fourier descriptors with applications to objects recognition in SVM context. *Journal of Mathematical Imaging and Vision*, 30(1):43–71, 2008.

[179] http://ip.hhi.de/imagecom_G1/savce/downloads/. JSVM Reference Software. Version 8.6.

[180] G. Sreelekha and P. Sathidevi. An improved jpeg compression scheme using human visual system model. *Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*, pages 98–101, June 2007.

[181] M. Stojmenovic, A. Solis-Montero, and A. Nayak. Colour and texture based pyramidal image segmentation. In *2010 International Conference on Audio Language and Image Processing (ICALIP)*, pages 778–786. IEEE, Nov. 2010.

[182] C. Strauss. *Low complexity methods for interpolation and pseudo semantic extraction: applications in the LAR codec*. PhD thesis, INSA Rennes, November 2011.

[183] C. Strauss, F. Pasteau, F. Autrusseau, M. Babel, L. Bédat, and O. Déforges. Subjective and Objective Quality Evaluation of LAR coded Art Images. In *Proceedings of IEEE International Conference on Multimedia and Expo, ICME'09*, pages 1–4, Cancun, Mexique, July 2009.

[184] P. Strobach. Tree-Structured Scene Adaptive Coder. *IEEE Trans. on Communication*, 38(4):477–486, April 1990.

[185] T. Strutz and E. Müller. Image data compression with pdf-adaptive reconstruction of wavelet coefficients. In *Proceedings of SPIE - The International Society for Optical Engineering*, pages 747–758, 1995.

[186] G. J. Sullivan and R. L. Baker. Motion compensation for video compression using control grid interpolation. In *Proceedings of International Conference on Speech and Signal Processing*, pages 2713–2716. IEEE, 1991.

[187] K. Tahboub and H. Asada. A compliant semi-autonomous reactive control architecture applied to robotic holonomic wheelchairs. In *Advanced Intelligent Mechatronics, 1999. Proceedings. 1999 IEEE/ASME International Conference on*, pages 665 –670, 1999.

[188] D. Taubman and A. Zakhor. Multirate 3-d subband coding of video. *IEEE Transactions on Image Processing*, 3(5):572–588, Sep 1994.

[189] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[190] D. M. Thomas. A study on block matching algorithms and gradient based method for motion estimation in video compression. In D. Nagamalai, E. Renault, and M. Dhanuskodi, editors, *Advances in Digital Image Processing and Information Technology*, volume 205 of *Communications in Computer and Information Science*, pages 136–145. Springer Berlin Heidelberg, 2011.

[191] J. Tian and R. O. Wells. Reversible data-embedding with a hierarchical structure. In *ICIP*, volume 5, pages 3419–3422, Oct. 2004.

[192] A. M. Tourapis. Enhanced predictive zonal search for single and multiple frame motion estimation. In *SPIE VCIP'02*, pages 1069–1079, 2002.

[193] A. M. Tourapis. JM reference software version 14.0. Doc. JVT-AE010, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, London, UK, 2009.

[194] D. Tschumperle and R. Deriche. Vector-valued image regularization with pdes: A common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:506–517, 2005.

[195] M. Urvoy. *Les tubes de mouvement : nouvelle représentation pour les séquences d'images (Motion tubes: a new representation for image sequences)*. PhD thesis, INSA Rennes, March 2011.

[196] M. Urvoy, N. Cammas, S. Pateux, O. Déforges, M. Babel, and M. Pressigout. Motion tubes for the representation of images sequences. In *Proceedings of ICME'09*, pages 1–4, Cancun, Mexique, July 2009.

[197] S. R. Vantaram, E. Saber, S. Dianat, M. Shaw, and R. Bhaskar. An adaptive and progressive approach for efficient gradient-based multiresolution color image segmentation. In *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pages 2345–2348, Piscataway, NJ, USA, 2009. IEEE Press.

[198] A. Wang, Z. Xiong, P. A. Chou, and S. Mehrotra. Three-dimensional wavelet coding of video with global motion compensation. In *Proceedings of Data Compression Conference (DCM'99)*, pages 404–413, 1999.

[199] Y. Wang, S. Rane, P. Boufounos, and A. Vetro. Distributed Compression of Zerotrees of Wavelet Coefficients. In *IEEE International Conference on Image Processing (ICIP)*, September 2011.

[200] A. Watson, G. Yang, J. Solomon, and J. Villasenor. Visibility of wavelet quantization noise. *Image Processing, IEEE Transactions on*, 6(8):1164–1175, 1997.

[201] L.-Y. Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of ACM SIGGRAPH*, pages 479–488, New York, USA, 2000.

[202] M. Weinberger, G. Seroussi, and G. Sapiro. Loco-i : a low complexity, context-based, lossless image compression algorithm. *Proc. Of the IEEE Data Compression Conference*, pages 141–150, 1996.

[203] M. J. Weinberger, G. Seroussi, and G. Sapiro. LOCO-I: A low complexity, context-based, lossless image compression algorithm. In *Proc. on Data Compression conference*, pages 140–149, Snowbird, UT, March 1996.

[204] W. Wharton. *Principles of television reception*. Pitman, London, 1971.

[205] T. Wiegand and H. Schwarz. *Source Coding: Part I of Fundamentals of Source and Video Coding, Foundations and Trends in Signal Processing*, volume 4. Foundations and Trends in Signal Processing, 2011.

[206] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *Circuits and Systems for Video Tech., IEEE Trans. on*, 13(7):560–576, July 2003.

[207] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien. Joint Draft 10 of SVC amendment. Doc. JVT-W201, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, San Jose, CA, USA, Apr. 2007.

[208] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien. WD2: Working Draft 2 of High-Efficiency Video Coding. Doc. JCTVC-D503, Joint Collaborative Team on Video Coding (JCT-VC) ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, San Jose, CA, USA, Jan. 2011.

[209] P. Willemin, T. Reed, and M. Kunt. Image Sequence Coding by Split and Merge. *IEEE Trans. on Communication*, 39(12):1845–1855, December 1991.

[210] J. Wolf, A. Wyner, and J. Ziv. Source coding for multiple description. *Bell System Technical Journal*, 59(8):1417–1426, 1980.

[211] X. Wu. Lossless Compression of Continuous-Tone Images, via Context Selection, Quantization and Modelling. *IEEE Trans. on Image Processing*, 6(5):656–664, May 1997.

[212] X. Wu, G. Zhai, X. Yang, and W. Zhang. Adaptive sequential prediction of multidimensional signals with applications to lossless image coding. *Image Processing, IEEE Transactions on*, 20(1):36 –42, jan. 2011.

[213] Z. Xiong, X. Sun, and F. Wu. Block-based image compression with parameter-assistant inpainting. *Image Processing, IEEE Transactions on*, 19(6):1651 –1657, June 2010.

[214] L. Xu, S. Ma, D. Zhao, and W. Gao. Rate control for scalable video model. In *Visual Communications and Image Processing.*, volume 5960, pages 525–534, 2005.

[215] X. Yang and K. Ramchandran. A low-complexity region-based video coder using backward morphological motion field segmentation. In *IEEE Transactions on Image Processing*, volume 8, pages 332–345, March 1999.

[216] G. S. Yovanof and S. Liu. Statistical analysis of the DCT coefficients and their quantization error. In *1996 Conference Record of the Thirtieth Asilomar Conference on Signals, Systems and Computers, 1996*, volume 1, pages 601–605, Nov. 1996.

[217] L. Yuan Wu and Z. Shouxun. Optimum bit allocation and rate control for H.264/AVC. Technical report, Joint Video Team, doc. JVT-O016, 2005.

[218] L. L. Yz and S. K. M. Y. Color Image Segmentation: A State-of-the-Art Survey. *Proceedings of the Indian National Science Academy*, 67(2):207–221, 2001.

[219] X. Zhao, J. Sun, S. Ma, and W. Gao. Novel statistical modeling, analysis and implementation of rate-distortion estimation for h.264/avc coders. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(5):647–660, May 2010.

[220] C. Zhu, X. Sun, F. Wu, and H. Li. Video coding with spatio-temporal texture synthesis. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 112 –115, july 2007.

From image coding and representation to robotic vision

Abstract. This habilitation thesis is first devoted to applications related to image representation and coding. If the image and video coding community has been traditionally focused on coding standardization processes, advanced services and functionalities have been designed in particular to match content delivery system requirements. In this sense, the complete transmission chain of encoded images has now to be considered.

To characterize the ability of any communication network to insure end-to-end quality, the notion of Quality of Service (QoS) has been introduced. First defined by the ITU-T as the set of technologies aiming at the degree of satisfaction of a user of the service, QoS is rather now restricted to solutions designed for monitoring and improving network performance parameters. However, end users are usually not bothered by pure technical performances but are more concerned about their ability to experience the desired content. In fact, QoS addresses network quality issues and provides indicators such as jittering, bandwidth, loss rate...

An emerging research area is then focused on the notion of Quality of Experience (QoE, also abbreviated as QoX), that describes the quality perceived by end users. Within this context, QoE faces the challenge of predicting the behaviour of any end users. When considering encoded images, many technical solutions can considerably enhance the end user experience, both in terms of services and functionalities, as well as in terms of final image quality. Ensuring the effective transport of data, maintaining security while obtaining the desired end quality remain key issues for video coding and streaming.

First parts of my work are then to be seen within this joint QoS/QoE context. From efficient coding frameworks, additional generic functionalities and services such as scalability, advanced entropy coders, content protection, error resilience, image quality enhancement have been proposed.

Related to advanced QoE services, such as Region of Interest definition of object tracking and recognition, we further closely studied pseudo-semantic representation. First designed toward coding purposes, these representations aim at exploiting textural spatial redundancies at region level.

Indeed, research, for the past 30 years, provided numerous decorrelation tools that reduce the amount of redundancies across both spatial and temporal dimensions in image sequences. To this day, the classical video compression paradigm locally splits the images into blocks of pixels, and processes the temporal axis on a frame by frame basis, without any obvious continuity. Despite very high compression performances such as AVC and forthcoming HEVC standards , one may still advocate the use of alternative approaches. Disruptive solutions have also been proposed, and offer notably the ability to continuously process the temporal axis. However, they often rely on complex tools (*e.g.* Wavelets, control grids) whose use is rather delicate in practice.

We then investigate the viability of alternative representations that embed features of both classical and disruptive approaches. The objective is to exhibit the temporal persistence of the textural information, through a time-continuous description.

At last, from this pseudo-semantic level of representation, texture tracking system up to object tracking can be designed. From this technical solution, 3D object tracking is a logical outcome, in particular when considering vision robotic issues.